

# What future for Europe's universities?

**The Maastricht Treaty (if ratified) would give the European Communities coordinating powers in education: what hope is there that the outcome will be a better system of universities?**

WHAT is to happen to the universities of Europe? Paradoxically, closer economic integration of Europe has been accompanied by what seems to be an even more emphatic national or even regional devolution of university affairs. That is not surprising. One of the promised benefits of the European enterprise is that it will allow distinctive regional cultures to flourish. One federalist view of Europe's future is that of a "Europe of the regions", in which German *Länder*, Italian pre-Garibaldi regions and even the ancient kingdoms or provinces of Britain (Scotland, Wales and even Ulster) would separately become equipotent members of a federal state. Would it not be natural that they should also run culturally distinctive universities?

That is one theory, not especially compelling, which does not fully account for the devolutionary tendencies of the past few years. It may be more relevant that hopes by central governments, embarrassed by the costs of more liberal access to higher education, that regional enthusiasm might bring savings have inevitably surfaced. This partly explains the snap decision of the British government that there should be separate arrangements for supporting higher education in Scotland and in Wales. (In Germany, on the other hand, the constitution confirms the tradition of centuries that the *Länder* governments are the prime supporters of their universities.) There is substance in the view that universities commanding regional loyalty are the stronger on that account. Moreover, there is ample evidence that parochial connections do not prevent universities from acquiring international reputations (witness several campuses of the University of California, not to mention Heidelberg.) But there are also dangers.

## Dangers

One of them is that the Maastricht Treaty, if ratified in some form (but renegotiation would be better), will give the European Communities (EC) competence in the field of education. Although the treaty prevents EC interference in member states' existing curricula or forms of organization, it will be able, if a majority of members agrees, to supplement what national governments are already doing. That new freedom could either be an opportunity or the opposite, the makings of a disaster. It

is not too soon for academics and researchers to be telling the EC what they expect of it.

Here are some essential ingredients of a shopping list for Brussels. Although most member governments now laudably accept ambitious goals for the participation of young people in higher education, they often do so by regarding their universities as machines for performing that service. Too many European universities, in other words, are less in charge of their own affairs than they should be. They must take in the students who choose to present themselves. They must teach what the government, as proxy for the students, requires. And those who teach are usually employees of the government, not of the institutions at which they work. In most places, politeness has engendered a seemly recognition that academics, unlike mailmen for example, must enjoy some degree of independence. But the principle that successful universities are self-governing institutions, determining for themselves what is taught by whom to whom, is formally a dead letter in Europe. That issue, sadly, is beyond the scope of the EC.

## Mobility

Another issue needing attention is that of mobility. Already it has been agreed that students can be educated where they choose within the EC, paying fees no different from those expected of domestic students. But teachers are not so well dealt with. Where academics are technically civil servants, foreigners may be denied permanent membership of a university faculty. (France is an honoured exception.) The good reason why that rule needs changing is that only thoroughly cosmopolitan universities (Harvard, for example) have been shown by experience to be capable of outstanding excellence. Parochialism is an impediment to Europe's emulation.

More humdrum considerations work against the beneficent notion of a thoroughly mobile force of European academics — salaries, for example. Tempting a German academic to a British university is virtually impossible; to a Danish university only a little less difficult. Everybody acknowledges that Swiss academics are virtually unemployable elsewhere. The task of removing these disparities, organizational as they are, will not fall within the EC's new competence, which may be just as well:



Europe's university system needs to be affordable as well as excellent. But there is scope for EC help with the mobility of academics towards universities judged to have excellence within their grasp.

How should this be done? The temptation to single out in advance a group of European universities for special treatment should be resisted. The result would be a series of political dogfights more intense than those provoked by last week's settlement of the row with the United States over oilseed production. What the EC needs, instead, is a scheme for awarding grants to institutions for academic innovations (often research projects) judged worthwhile by serious panels of referees and which, on that account, command the respect of the international community. That way, over time, the outstanding universities would emerge head and shoulders above the merely excellent. To lend piquancy to the recipe, grant-making should be extended not merely to candidate members of the EC, but to the universities of Poland, Czechoslovakia and Hungary, traditionally part of the European academic circus.

Sadly, this important issue is unlikely to be raised at the Edinburgh meeting of Europe's Council of Ministers two weeks from now. Indeed, on past EC form, it is unlikely ever to rate a well-informed discussion. That is the outcome most urgently to be avoided. Europe, nearly a millennium ago, invented the idea that groups of scholars might educate the young in their ways. (At Bologna, at the beginning, members of the faculty were appointed and paid by their students.) It would be a great pity if Europe now let slip the opportunity to make its universities its agents for the preservation of the old culture, for sometimes subversive social change and for creating the prosperity it repeatedly promises. □

## His father's Oldsmobile

**The heads of the NSF and NIH in the United States have followed ways of plotting the future.**

WITH governments everywhere eager to harness science to economic development, last week's report on the future of the US National Science Foundation (NSF) is a timely warning of the limits of what is possible (see page 285). But the report also reveals important differences in style and substance between Walter Massey, the NSF director, and Bernadine Healy, the director of the US National Institutes of Health (NIH), who has been engaged in a similar review of her agency's future.

Both Massey and Healy, appointed well into the term of outgoing president George Bush and waiting to learn whether they will be replaced by president-elect Bill Clinton, believe strongly that their agencies should do more than serve the needs of their constituent popula-

tions of scientists. But while Healy has spent 18 months on a highly publicized and at times controversial effort to produce a strategic plan that she has labelled *Advantage America*, Massey has achieved much the same with a minimum of fuss and without drawing unfavourable attention to himself or his agency.

Part of the difference lies in personalities: Healy seems to relish confrontation, Massey avoids it whenever possible. Thus when the director of the National Aeronautics and Space Administration told Congress that the space station Freedom would contribute significantly to basic biology by studying how astronauts function under microgravity, Healy replied immediately that more could be learned by increasing NIH's budget to study such processes on Earth. In contrast, when the US Senate told NSF to be more responsive to the needs of industry, Massey followed the prudent course of appointing a 15-member commission to study the issue.

Healy is fond of saying that business-as-usual is no longer good enough and that NIH must change. She has even adapted a popular commercial for Oldsmobile motor-cars that emphasizes how the automobile has been redesigned to appeal to a younger audience, telling researchers around the United States that "this is not your father's NIH". But even when biomedical researchers agree that change is needed, they resent being told that they are part of the problem, and they are confused by Healy's incessant criticism of a system that they feel has worked rather well for four decades. At the same time, there is grumbling within Congress that the strategic plan, begun shortly after Healy took office in April 1991, is languishing within the Department of Health and Human Services.

In contrast, Massey says that he is quite pleased with the commission's conclusion that there is nothing wrong with NSF and that no drastic changes are necessary. In other words, the agency is still his father's NSF, and he likes it that way. But he also agrees that NSF can work more closely with industry and he promises to increase industrial collaboration in the months ahead. His response is intended both to assure researchers that he is on their side and to tell Congress that he plans to listen carefully to its advice. Any disagreement can be ironed out, he says, especially if Congress sees fit to increase NSF's budget.

Clinton has not yet said whether he plans to replace either director — both serve at the pleasure of the president, although Massey officially has a six-year term of office that runs until 1996. In the end, the decision may come down to personal considerations. It would be wrong if the choice of the head of a strictly science agency was determined by political philosophy. Still, if one remains and one is sent packing, it may be well to remember that each handled differently their agency's attempt to plot its future. □

# NSF gets a pat on the back in report suggesting stronger industrial ties

**Washington.** US academic scientists can take comfort from a report issued last week that says the National Science Foundation (NSF) is not broken and does not need to be fixed by shifting more of its budget into applied research. The report, by a Commission on the Future of NSF, applauds everything that NSF is doing with its \$2-billion research budget and urges NSF officials to work harder at explaining how basic research eventually benefits society.

Many were troubled when Walter Massey, the NSF director, announced in August that the National Science Board was forming a 15-member commission of prominent academic and business leaders to examine how NSF might play a larger role in strengthening the US economy (see *Nature* **358**, 611; 1992). Coming on the heels of a suggestion from a committee of the US Congress that NSF should help the country to regain economic preeminence, there was concern that the foundation might begin making large grants to industrial scientists and reduce its traditional emphasis on investigator-initiated basic research. Despite three public meetings during which members went to great lengths to disavow such intentions (see *Nature* **359**, 261; 1992), many were waiting anxiously to read the report itself.

What they will find is 11 pages of praise for NSF's current activities and its mission to support academic research and improve science and mathematics education. To be sure, there is room for improvement: NSF should support more interdisciplinary research, work more closely with industry and evaluate more carefully the outcomes of research it funds. NSF's priorities should be linked to a broader national mission, the report emphasizes, and NSF's advisory body, the National Science Board (NSB), should play a larger role in setting those targets. But the report emphasizes that science is not to blame for the difficulties facing US industry, saying that "failures in the market place have not been the result of a slow transfer of academic science to industry".

The report is likely to placate critics in Congress concerned that NSF is not doing enough to help the country without jeopardizing its role as a primary source of funding for academic researchers outside the biomedical sciences. Stanford chemist Richard Zare, a NSB member, ruled out any change in the status quo by asking if the report proposes direct funding of industrial research and being told by William Danforth, co-chair of the commission and chancellor of Washington University in St Louis, Missouri, that "it's not on anybody's agenda".

Another important audience for the

report is industrial researchers and corporate leaders. The commission's other co-chair, Robert Galvin, the former head of Motorola Inc., says that the three-month effort opened his eyes to NSF's value as an untapped resource for industry.

As an example, Galvin draws attention to a recommendation by the commission that more industrial scientists should serve on NSF advisory committees and accept temporary assignments at the foundation. Such 'rotators' are predominantly from universities, however, and Galvin confessed that it had never occurred to him that one of the company's scientists might work for a few years at NSF. The report has changed his thinking, says Galvin. Similarly, Galvin hopes the report will convince other industrial leaders to pay closer attention to what NSF is doing and to provide matching funds for collaborative research.

Although the report does not contain numbers, it says that its proposals will require a larger NSF budget. If the federal government cannot provide such an increase, NSF should reallocate existing programmes to reflect national needs. At the same time, it suggests that NSF should seek more money from industry "to complement public funding for selected science, engineering and technology programs".

The report avoids discussing what NSF's research priorities should be, with Galvin and Danforth emphasizing that the commission lacked sufficient time and expertise to do so. That is an important difference from the 18-month exercise being conducted by the National Institutes of Health, which has compiled several hundred pages listing which areas of biomedical research are most deserving of support. That exercise, including five public hearings around the country, has touched off loud debates; in contrast, the NSB commission says that such details are best handled by NSF officials in the course of their work.

The commission, having delivered its report on time, will be dissolved and its findings will form the basis of discussion within the science board over the next several months. Although Congress has asked Massey to incorporate relevant portions of the report into an operating plan for 1993 that he must submit on 15 December, it is doubtful whether its recommendations will significantly alter the foundation's present direction. A more likely target is the foundation's budget for the 1994 fiscal year beginning on 1 October 1993, which will emerge in February as part of the first budget submitted by President-elect Bill Clinton.

**Jeffrey Mervis**

## Sakmann starts German-Israeli fund

Munich. Last year's Nobel prizewinner for physiology and medicine has donated much of his prize money to create an annual lecture tour for young scientists in Israel and Germany. The gesture is one more sign of

increasing scientific cooperation between the two countries.



**Bert Sakmann**

For German electrophysiologist Bert Sakmann, who shared the 1991 Nobel Prize with Ernst Neher for their invention of the patch-clamp technique, the DM150,000 (US\$100,000) endowment was an

opportunity to reinforce links to the past. "We still want to do something for Israel, especially for young scientists", he says.

The lectureship is named after his mentor Bernard Katz, a physiologist forced to leave

Germany in the 1930s who established strong scientific links — particularly in cell physiology — with Israel after the war. Sakmann studied with Katz at University College London.

A second inspiration was Wilhelm Feldberg, a physiologist who was also forced to emigrate from antisemitic prewar Germany. Feldberg used the compensation given to him by the German government to found the Feldberg lecture series based in Britain.

Relations between German and Israeli scientists have been improving since the early 1960s, when Germany's Max Planck Society and Israel's Weizmann Institute established the Minerva Foundation to administer a joint programme of research. The German-Israeli Foundation (GIF) was established in 1988 with DM150 million contributed by the two governments.

The Bernard Katz lectureship is open to young German and Israeli cell physiologists and biophysicists. The award, which will alternate between the two fields, provides money for them to visit colleagues in the other country.

**Alison Abbott**



# Germany will ease requirements of gene technology laws in bow to researchers

**Munich.** Germany's federal government has agreed in principle to relax the way its gene technology laws are administered after a long campaign by scientists to reduce the number of forms and lengthy authorization procedures.

But the changes come too late for a laboratory in Marburg, which was closed down last week by the regional Ministry of the Environment. Researchers, whose equipment and notebooks have been seized even though the research had previously been designated as without risk, believe the action was an attempt to forestall liberalization of the laws.

Molecular biology in Germany has been undermined by a powerful anti-genetic-engineering lobby that has stifled research and left the country trailing the rest of Europe. Laws introduced in July 1990 continued the trend by requiring elaborate safety procedures and time-consuming recordkeeping — irrespective of the level of risk involved in experiments (see *Nature* 359, 93; 1992). But a concerted campaign by academic and industrial scientists has now borne fruit.

The government has agreed that procedures for monitoring experiments will be relaxed. Gebhard Ziller, state secretary at the ministry of research and technology, last Friday told a meeting on gene technology organized by the European Patent Office in Munich that the government would be "taking steps to remedy the morass of bureaucratic procedures" by amending the Genetic Engineering Act and simplifying the administration. The changes would not lower safety standards, he

assured critics. Experiments that by definition pose no risk will no longer need to be reported to the authorities once a general authorization has been received. The government is also expected to remove an automatic three-month delay in granting permission for experiments designated as 'very low risk'. The law is on the second of three readings in parliament, and the

compliance with the gene laws found that a project run by Klaus Hazemann, head of haematology and oncology at a local hospital, had been moved to an identical laboratory in another part of the building.

Two weeks later, the local government decided to take action. Last Thursday, police seized refrigerators and freezers to check the plasmids being used — an almost impos-

sible task estimated to involve the sequencing of up to half a billion kilobase pairs. They also took notebooks that did not conform to the strict requirements of the law. A spokesperson for the ministry said that the law had been broken and that researchers must start the complex application procedure from scratch.

Response has verged on the hysterical. Local papers have reported that the laboratory was releasing 'carcinogenic bacteria' into the environment. In fact, Hazemann is looking at IGF binding proteins that have a negative effect on cellular proliferation and is investigating the mechanisms controlling the genes for these proteins. The work, which has implications for some types of lung cancer, had previously been defined as belonging to the

lowest of the four categories of risk level outlined in the gene law, that is, posing no risk for human health or the environment, although the laboratory was originally registered for second level — very low-risk — experiments.

Hans-Dieter Klenk, a member of the university safety commission with responsibility for the institute's laboratories, says that the research group erred in not reporting the change in location, but he disputes claims that the new room is unsafe. Everything was in order during his own inspection, he said.

Klenk says that the proposed changes in the gene laws are extremely important for Hessen, where an anti-gene technology government has hindered research by enforcing the letter of the law. It is not uncommon for scientists to wait a year for permission to conduct a simple experiment.

Researchers throughout Germany welcome the changes. Ernst Winnacker, head of the Biochemistry Institute in Munich, says 80 per cent of all experiments using gene technology are designated 'no risk', and a further 17 per cent are 'very low risk'. "But most important", says Winnacker, "it's a signal that [German science] laws can indeed be adapted to technological changes".

**Alison Abbott**



**Protesters outside last week's meeting in Munich at the European Patent Office display signs reading, "Gene technology — the ultimate exploitation of all living things".**

Ministry of Public Health is confident that it will be in place by January 1994.

In the meantime, the Department of the Environment in Hessen, one of Germany's 16 federal states and led by the Green party, has unexpectedly closed a research laboratory at the Institute of Molecular Biology and Tumour Research of the University of Marburg. Health and safety officers conducting a routine inspection to check

## Protesters picket patent office

**Munich.** The European Patent Office has become the target of anti-genetic-engineering activists after its recent granting of Europe's first animal patent — the Harvard oncomouse. Although last Friday's symposium (see above), attended by both scientists and representatives of lobbying organizations, was intended to ease public concern over genetic engineering, the local group *Kein Patent auf Leben* (No Patents on Life) claimed that its opinion has continued to be ignored.

As a consequence, the group assembled 150 demonstrators from five European countries to picket the meeting. A spokesman, Christof Then, complained that there was no opportunity for objectors to lobby participants directly and promised that his group would continue protesting against the granting of patents on life forms both in Germany and at the European Communities headquarters in Brussels.

"We are against biotechnology in general", he says. "We are against relaxing of the gene laws; with such a young technology it's too dangerous."

**A.A.**



# Freer trade in US oranges helps Japanese researchers

**Fukuoka, Japan.** Pressure to open up Japan's orange market to US farmers is hurting native growers but helping university researchers. Administrators of Kyushu University in Fukuoka city in Japan's southern island of Kyushu have begun drafting plans to move the university's main

Kyushu University is inconveniently divided among four campuses in Fukuoka. Under the plan now being drawn up, the main Hakozaki campus near the centre of Fukuoka that houses the faculties of science, engineering, agriculture, literature, economics and education will move to the new site in the far western outskirts of Fukuoka, joining the college of general education now on a separate campus in the southwestern part of the city. The site has become available because unprofitable *mikan* farms are being sold following a recent US-Japanese trade agreement that has opened up Japan to a flood of US oranges.

The university hopes to receive twice as much space in return for its prime real estate in the centre of Fukuoka, according to the dean of one of the university's faculties, "with perhaps a few buildings included as well". But the site is hilly and will need levelling before the move, expected to be in about five years. The plan must first be approved by the Ministry of Education, Science and Culture and the finance ministry, but it seems likely to go ahead because it supports a government policy to improve the environment of Japan's national universities.

Tokyo University has similar plans to expand into the countryside. Next fiscal year, the university hopes to receive money for detailed plans to move several parts of the univer-

sity from central Tokyo to Kashiwa, northeast of the capital in Chiba Prefecture. Among the plots of land the university hopes to sell is that occupied by the Institute of Solid State Physics in Roppongi, a popular night spot. Parts of the faculties of engineering and science also plan to move to the Kashiwa campus.

The first national university expected to sell land under the new amendment is Osaka University, which has already moved from central Osaka to Suita campus in the northern outskirts of the city. However, these plans may be affected by the recent bursting of Japan's 'bubble economy' which has sent land prices plummeting, particularly in the centre of cities. One of the deans of Tokyo University doubts that the university will be able to find a buyer in the present depressed economic climate. Fortunately, the new law allows the government to obtain loans for the relocations until the real estate market recovers.

David Swinbanks

## Reforms accompany relocations

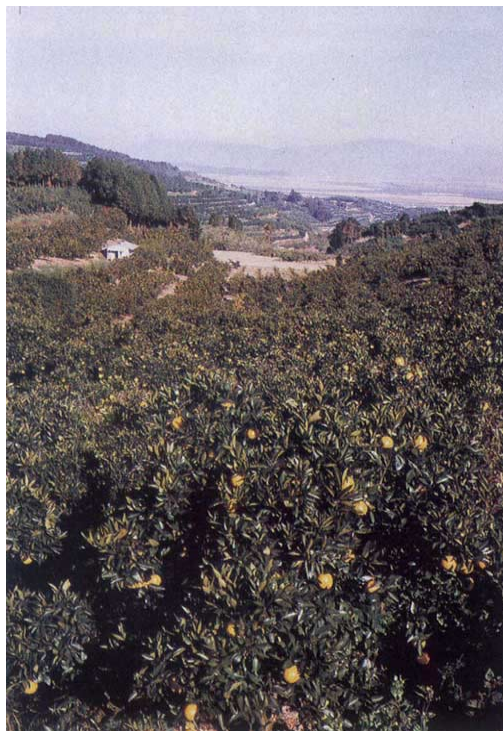
**Fukuoka.** As well as moving from the city to orange fields (see above), Kyushu University, like many of its counterparts elsewhere in Japan, will undergo major reform and reorganization as it tries to reinvigorate its research environment.

The college of general education, which provides two years of initial training for all students, will be dissolved. Some of the general education faculty members will move to a new graduate school while others will join faculties on the main campus.

There is a trend among Japan's national universities to eliminate the unpopular general education courses established during the US occupation of Japan after the Second World War. The general education classes are overcrowded and are often housed in cramped quarters. The quality of teaching is often low, and faculty have little time for research.

Large, multidisciplinary research groups (*multi-koza*) will be formed within the science and engineering graduate schools of Kyushu University instead of the present small groups (*koza*) run by one professor. The idea is to make the graduate research system more flexible and to encourage new fields of research; it is modelled on an initiative by Tokyo University (see *Nature* **351**, 679; 1992).

D.S.



Trade policies peel off *mikan* farmers.

campus to a large plot in the outskirts of the city being vacated by mandarin orange (*mikan*) farmers put out of business by the recent liberalization of Japan's orange market.

Kyushu University is just one of Japan's leading national universities taking advantage of amended laws allowing them to use money from the sale of university land to buy land for new university campuses. The amendment, effective from July this year, is one of several changes aimed at renovating Japan's dilapidated and cramped national universities. Government neglect has created the problem, made worse by a rapid decline in the student-age population beginning this year which is forcing universities to become more competitive.

Before the amendment, money from the sale of university land was deposited in the coffers of the Ministry of Finance and could not be designated for university use. But the money will now be pooled in a fund that universities can tap for buying new land.

## India threatened with loss of nuclear fuel

**New Delhi.** The French government has threatened to stop supplying fuel for India's two nuclear power plants at Tarapur next year unless India signs the Nuclear Non-Proliferation Treaty (NPT) or allows international inspection of its facilities.

The French have supplied the Tarapur plants with enriched uranium fuel since 1983, three years after the United States prematurely ended an agreement to supply them with fuel for their lifetime, but the French agreement ends in 1993.

Although France had not signed the NPT when it began its shipments, it became a member last year. In recent months, France has told India that it cannot continue the shipments under the current agreement. India considers the NPT to be discriminatory and it refuses to allow inspections by the International Atomic Energy Agency, claiming them to be an infringement of sovereignty.

K.S. Jayaraman

# Clinton plans offer hints of technology agenda

**Washington.** Ever since his election victory earlier this month, US President-elect Bill Clinton has uttered hardly a word about future policies, including those involving science and technology. However, this silence has not dampened speculation that Clinton's enthusiasm for federal research and development is more likely to pay off for technology and applied research than it is for basic science.

One clue to Clinton's thinking are the criteria for selecting the president's science adviser, who doubles as director of the White House Office of Science and Technology Policy. Clinton is looking for someone with strong industrial experience and high standing in the academic community, according to an aide to Senator Al Gore, the Vice President-elect. The aide, who has already collected the names of more than a hundred possible candidates for the position, offers as an example the résumé of chemist Mary Good, vice-president for technology at Allied-Signal Inc. and a former chair of the board that oversees the National Science Foundation. Good may not be offered the job, but it is clear that her background is preferable to that of physicist D. Allan Bromley, the current science adviser who has spent most of his career at Yale University and has never worked in industry.

What this means for Clinton's policies towards research depends on the authority given to the science adviser, of course. Traditionally, the science advisor has been relegated to coordinating federal research

policy after the fact or, worse, writing reports and convening committees to give the appearance of such a policy where none exists. But the presence of Gore, arguably the most science-literate politician ever to reach the White House, bodes well. Clinton has designated Gore as his technology tsar, using research to strengthen US industry in the same way that Dan Quayle, the outgoing vice president, tried to use regulatory reform. Although the new science adviser may report directly to Gore rather than to Clinton, a mandate from the White House is seen as preferable to Bromley's ostensible access to the president.

Another telling indicator is the composition of Clinton's economic transition team. Voters elected Clinton because of their concern about the economy, and he sees a strong research base as essential for an industrial recovery. His choice of Robert Reich, a Harvard University economist and advocate of industrial policy to lead the economic transition team reflects this agenda. In turn, Reich has selected Laura D'Andrea Tyson, an economist at the University of California, Berkeley, and another proponent of industrial policy, to guide a team looking at technology and manufacturing policy.

Their mission is to translate Clinton's promises into concrete policies for the new administration. This includes helping to select people for important positions, providing details of how Clinton would increase domestic research spending to at least match military spending and tilt government-

sponsored research towards industrial needs. Tyson, a coauthor of a report entitled *Linking Trade and Technology Policies* released earlier this year by the US National Academy of Engineering, believes that government can help US industry compete without erecting trade barriers.

Finally, there are the opinions of Clinton's campaign advisers, whose stature ranges from having provided one-on-one briefings to having simply sent a letter to the candidate. One genuine inside player is the Council on Competitiveness, a private think-tank set up by John Young, the former head of Hewlett-Packard and a prominent supporter during the campaign. Last week, at a seminar on science policy under the new administration sponsored by the American Association for the Advancement of Science, the executive vice president of the council, Daniel Burton, predicted that the Clinton presidency would include:

- Greater effort to involve the private sector in setting US priorities for research and development;

- More government centres, perhaps under the auspices of the National Institute of Standards and Technology, for collaborative research with industry and to commercialize government-funded technology;

- More investment in "technology infrastructure", a term that is somewhat undefined but is believed to include high-performance computing and networks;

- A greater focus on manufacturing processes. During the campaign, Clinton promised to create 170 manufacturing extension centres that would do for high-technology industries what agricultural extension centres have done for farming;

- A broader interest in global science, both to collect information and to harmonize policies. As Burton put it, "we have 50,000 troops in Japan and only five commercial officers. Something about that balance is wrong."

Much of this policy is expected to be delegated to the Department of Commerce, which has had a Technology Administration since 1989. Although the programme has been nearly invisible during the Bush administration, Clinton is said to favour giving it sufficient authority and money to carry out its mandate.

Basic research is not expected to be at the centre of whatever technology policies Clinton adopts. Yet his willingness to create new applied research programmes suggests that he does not intend to cannibalize basic research to promote his technology plans. After the concerns raised about the direction in which the National Science Foundation and the National Institutes of Health appear headed (see page 285), that may be welcome news to bench scientists.

**Christopher Anderson**

## EMF report draws fire

**Washington.** When the health effects of electric and magnetic fields (EMF) are studied, the only predictable conclusion is that the results will be controversial. In 1989, the US Environmental Protection Agency (EPA) started a furore when it released a report suggesting that EMF radiation might be linked to leukaemia and other cancers. Now, three years later, the other side has spoken.

A scientific panel commissioned by the Committee on Interagency Radiation Research and Policy Coordination, made up of representatives of several federal agencies, concluded in a report\* released last week that EMF "does not appear to constitute a public health problem" and that research on the subject should not receive more funding. But the ink was not even dry before the report was criticized by those who consider EMF radiation to be a significant health hazard.

Critics suggest that the 11-member panel, which reviewed about 1,000 published studies, was assembled primarily to rebut the

EPA report. They point out that one of the authors, Dimitrios Trichopoulos, a Harvard University epidemiologist, has testified on behalf of electric utility companies against any link between EMF and public health. (An official from Oak Ridge Associated Universities, which prepared the report, says that she was aware of Trichopoulos's role but did not feel that it presented a problem.)

Others point out that the report's executive summary, which recommends against a major expansion of a national EMF research programme, seems to run counter to its individual chapters, which urge dozens of new studies. In a letter to the panel sent on 30 October, William Farland, director of health and environmental assessment at EPA, asked that the panel make clear that the report does not reflect the views of the agencies that paid for it.

**Christopher Anderson**

\* *Health Effects of Low-Frequency Electric and Magnetic Fields*, Oak Ridge Associated Universities, 1992 (US Government Printing Office, publication no. 029-000-00443-9).



# FDA wants to invite women into drug trials earlier

**Washington.** In response to changing attitudes towards women, the US Food and Drug Administration (FDA) plans to loosen rules banning women of childbearing age from enrolling as subjects in the early tests of new drugs. The ban, which was prompted by fears that women in trials might get pregnant and bear deformed children, has been contested for several years by critics who say it violates women's civil rights. But even if FDA rescinds the ban, premenopausal women will probably continue to be ex-

cluded from early trials because of fears by companies that they might be sued if a fetus were damaged by the experimental drug.

IMAGE  
UNAVAILABLE  
FOR COPYRIGHT  
REASONS

## Many women want to be enrolled earlier.

cluded from early trials because of fears by companies that they might be sued if a fetus were damaged by the experimental drug.

FDA's regulations, formalized in 1977, began as unofficial guidelines in the 1960s after the thalidomide tragedy. The rules exclude women of childbearing age as subjects in drug trials until the drug concerned is known to be safe and effective. In practice, this rule eliminates women from phase I (toxicity) trials and the early part of phase II (small-scale efficacy) trials. Animal studies of the drug's effect on a fetus must also be completed before women enter trials. This provision effectively bars women from all but late phase II and phase III (large-scale efficacy) trials. The FDA does, however, waive these rules for drugs that could prolong or save lives.

There appear to be few compelling scientific reasons for including women in the earliest stages. Toxicity does not differ by gender; early enrolment would give investigators added time to work out the proper dosage and side effects in women, but the advantages are negligible. "There haven't been enough obvious [pharmacological] differences to make you think that women are different from men", says Leslie Benet

of the school of pharmacy at the University of California, San Francisco. The decision seems driven by political pressure and changing attitudes towards women. For the past several years, women within and outside FDA have complained that the ban is paternalistic and have urged the agency to allow women to choose their own level of risk exposure. The FDA is thought to apply a double standard; as evidence, women cite the recent trials for the prostate enlargement therapy Proscar, which at the time was thought to be a potential cause of birth defects. The trial's male subjects were simply asked to use contraception and trusted not to father any children.

As legal support for their position, critics of the ban point to a unanimous decision last year by the US Supreme Court that female workers at a lead battery factory could not be barred from jobs exposing them to substances harmful to fetuses. As a result, "you need very good reasons to keep women out", says Robert Temple of FDA's Center for Drug Evaluation and Research.

Having decided that the ban should be loosened, the FDA is developing rules giving women the choice to participate in early trials and protecting fetuses conceived by accident. Possible safeguards include a series of pregnancy tests during the trials and removing the ban only for drugs that leave the body too rapidly to affect a fetus. Some FDA officials favour revoking the ban altogether, as phase I trials consist of only a few dozen subjects taking a single dose of a drug.

No matter what FDA decides, concern that drug companies might be sued over a deformed fetus are likely to keep women of childbearing age out of early trials. Product-liability lawyers agree that even a woman who has signed an informed consent form might win a liability suit if damage to the fetus were severe enough. A suit could also be filed by an adult subjected *in utero* to such tests. Some companies have suggested that they would enrol women earlier if Congress passed a law shielding them from liability, but the prospects of such a law are poor.

Traci Watson

# NASA researchers fear crackdown after security lapse

**Washington.** Scientists at the National Aeronautics and Space Administration (NASA) Ames research centre in Mountain View, California, are concerned that academic freedoms may be restricted in the wake of an investigation by several federal agencies into a leak of secret computer codes and other abuses.

Last week, NASA issued a harshly worded press release blaming the "culture and environment" at Ames for a series of security breakdowns, some more than a year old. The scientists are "strongly biased towards maintaining an academic reputation, according to NASA, "rather than meeting US industry and national needs". As a result, NASA says that some aerospace companies are reluctant to share important data with NASA in case they are improperly disseminated. Daniel Goldin, the NASA administrator, called for a "culture change" at the laboratory, especially regarding contact with foreign researchers.

On the face of it, the NASA statement and a similar message given to Ames researchers at meetings last week imply that scientific and intellectual freedom at the laboratory may be curtailed. But those reading between the lines are playing down the impact on basic research. Most of the Ames scientists do applied work involving aerodynamic simulations and computer modelling, and the "security lapse" that triggered the investigation appears to have been a computer network leak of computer code for a hydrodynamic modelling algorithm. Those doing fundamental research — mostly in astronomy, cosmology and planetary studies — expect to continue interacting freely with outside colleagues.

The investigation involved some 30 NASA officials visiting the laboratory in August, locking researchers out of their offices while they looked for security leaks. Investigators found evidence of dozens of violations in security and procedures, ranging from an outside company operating within the laboratory to improper access to computers. The investigation appears to be part of a long-running conflict between Ames and NASA headquarters over the laboratory's independence rather than a response to a specific abuse.

Nevertheless, if NASA wants to bring Ames into line, life could become a good deal more difficult at the laboratory. Goldin has a reputation for forcing change and he may choose to make an example of Ames.

"A strong clampdown on exchanges does not bode well for our research mission", says Scott Sanford, an Ames physicist. "I hope the pendulum doesn't swing too far that way."

Christopher Anderson



# Researchers urge boycott of supply company in labour dispute with warehouse employees

**Washington.** The International Brotherhood of Teamsters Union has enlisted the support of scientists and physicians in a protracted labour dispute between Fisher Scientific of Pittsburgh, Pennsylvania, and 77 warehouse workers. The union is urging a boycott of products from Fisher Scientific until its members are rehired and their health-care coverage is restored.

The workers have been on strike since October 1991, when the company proposed a significant increase in the cost of their health-care insurance premiums. Union officials say that the company rejected a proposal by the union to take over the health plan, a step the union says would save the company at least \$600,000 over three years (the life of the worker's contract).

In February, the 77 workers were notified that they must accept the company's offer or lose their jobs permanently. The workers rejected the offer and, according to union officials, have since been locked out at the company's plant in Springfield, New Jersey. The plant serves as a major East Coast distribution point for Fisher Scientific, which has supplied the medical and scientific community with everything from chemicals and glassware to pH meters for 90 years.

In an open letter published in the 9 November issue of *The Scientist* newspaper,

the *ad hoc* Physicians and Scientists for Justice for Fisher Employees urged Fisher Scientific "to rehire the 77 Springfield workers and to reinstate their health insurance". Union representatives say that the group has 120 members.

The company disputes the union's description of the events surrounding the strike. A spokesperson, Harry Schwalb, says that Fisher Scientific considered the union's proposal to run its own health plan, but when more information was requested, none was forthcoming. He also said that Fisher Scientific proposed a "reasonable increase" in the insurance premiums, but nothing approaching a three-year increase of 450 per cent claimed by the union. In addition to the soaring cost of health insurance in the United States, Schwalb said that the increase reflected the "unusually high claims experience" at the Springfield plant. Although Fisher Scientific would not elaborate on what was "reasonable" or why claims at the plant were "unusually high", Schwalb said that the proposal would bring the premiums of the 77 Springfield workers in line with those paid by the company's other 3,100 employees.

The company also refuted union claims that it had replaced skilled workers with unskilled workers. A picker or packer of

orders is not required to have the same skills as the people who make electrochemical instrumentation or other precision equipment, Schwalb said.

Jonathan King, a professor of molecular biology at the Massachusetts Institute of Technology and a supporter of the campaign, hopes that his scientific colleagues will bring economic pressure to bear on the company by using alternative suppliers. "When a company whose profits come from a social concern for improved health decides to enhance its profit by cutting back on health insurance to a sector of its work force, then it's directly contradictory to the overall endeavour" [of providing the public with access to improved health care], King says.

In addition to involving researchers, the union has taken its case to sympathetic city officials. Last week, Boston City Council passed a resolution that would prohibit public institutions under the city's administration, including Boston City Hospital, from buying supplies from Fisher Scientific. The action awaits approval from the city's mayor, Raymond Flynn. Similar action is also planned for New York, the District of Columbia and Baltimore. "If we cut enough arteries, I believe we can win", says Ted Bloom, a union delegate. **Diane Gershon**

## Forthcoming space launch rekindles debate in Brazil

**São Paulo.** The scheduled launch next month by a US company of the first satellite built by Brazil has renewed debate over the direction of the Brazilian space programme. On the one hand, there is a feeling that the government has not done enough to support a civilian space programme and that it is more interested in saving money than in getting results. On the other hand, there is concern elsewhere that an indigenous capacity to launch future satellites could also be used to develop military weapons.

The Data Collecting Satellite (SCD-1), which arrived in the United States earlier this month, is being prepared for launch on 12 December from Florida atop the Pegasus rocket built by the Orbital Sciences Corporation of Fairfax, Virginia. The 50-foot winged vehicle is sent into orbit from an aircraft and can carry a payload of up to 450 kg. The 115-kg Brazilian satellite will be used to gather real-time meteorological data from automated ground stations and relay them to a control centre.

The Brazilian satellite will be only Pegasus's third launch. The rocket made its first flight in 1990, and its second launch

placed a US payload into the wrong orbit. Although Brazilians are relieved that a mission originally set for 1986 seems at last ready to go into orbit, some scientists are unhappy that the government did not choose a more tested rocket such as LTV's Scout, which last weekend successfully launched a military payload. It is thought that the low overall cost of the launch, US\$14 million, played a significant role in the government's decision.

At the same time, critics say that the government should have made a more favourable arrangement. José Monserrat Filho, an adviser on space policy to the Brazilian Society for the Progress of Science, says that the former Soviet Union was prepared to transfer space technology as part of an arrangement to use Kosmos, a Russian launch vehicle, and that an important opportunity to strengthen the country's rocket programme was lost by choosing a company that did not offer such benefits.

The original plan was to launch four small satellites built by the civilian National Institute of Space Research (INPE) and launched by a vehicle developed by the Air

Force. But the United States and other countries have raised objections for fear that the technology would also be applied to military payloads, in particular ballistic missiles (see *Nature* 339, 164 & 329; 1989). The strong ties between Brazilian weapons manufacturers and countries such as Iraq added to international concern.

The difficulties in obtaining the necessary equipment and technology have delayed the project, forcing the government to choose Pegasus for the SCD-1, which has been ready for more than a year. The satellite cost US\$20 million to build.

Although officials, including José Israel Vargas, the new minister for science and technology, smiled happily as the satellite was packed up for its trip, the fate of the Brazilian space programme remains uncertain. The new government of Itamar Franco, who replaced Fernando Collor de Mello as president after Collor's impeachment, has not yet taken a position on Collor's plan to create a single space agency. But most scientists would like to see it be placed under civilian control, as are the US and European space agencies. **Ricardo Bonalume**

# Armchair biodiversity

**SIR** — The word 'biodiversity' has become a popular symbol for arousing people's consciences about the extreme vulnerability of wild animals and plant populations. Thousands of pages have been devoted to equating 'biodiversity' with genetic diversity and interpreting it in terms of DNA. The problem of preserving biodiversity is frequently displayed as solely the task of collecting and storing the broadest number of different germplasms or DNA sequences to secure the future of species and the biological wealth of our world.

This armchair, molecular-biased approach is extremely dangerous, because real biodiversity is a consequence of diverse interactions between biological populations and environmental and anthropic factors. Wildlife preservation, the main component of biodiversity, has been successfully approached through monitoring of endangered species and by developing natural reserves and national parks. The problem becomes more complicated when synanthropic taxa are involved. Mankind has produced and driven biodiversity over thousands of years while interacting with domesticated plants and animals. Local cultivars cannot be simply stored without the understanding of the conscious and unconscious processes that led to their appearance and diffusion. Most of these processes ceased a long time ago in the so-called developed countries, with the transfer of these responsibilities from the local communities of peasants to specialized groups of technically qualified people. But cultivated plants and domesticated animals are so closely connected to the culture in which they arose that their survival is dependent on the survival of the culture. It is nothing new to hear about the extinction of cultivars or even cultivated species with the disappearance of the culture that raised them.

Extant aboriginal cultures, mainly illiterate, have been accumulating a considerable wealth of ecological knowledge and patterns of organisms adapted to their environments, and accordingly have produced a cohort of biodiversity in form of synanthropic plants and animals, which are now in a serious danger of loss because of the extensive and extremely intense acculturation imposed all over the world by unadaptive Western models.

Orally transmitted knowledge is extremely adaptive and dependent. Most of this information is understood only under the local environmental conditions by locally trained people, whereas Western efforts to collect in the short term the whole of aboriginal knowledge and patterns is biased by the importance

given to the taxon-use, which is explained exclusively in terms of active substances or prospective economic benefits.

The survival of both ecological and cultural diversity is the *sine qua non* for a long-term preservation of biodiversity. Germplasm can be preserved by means of freezing. But how can environmental or cultural diversity be frozen?

**Diego Rivera-Núñez**

**Concepción Obón-de-Castro**

*Departamento de Biología Vegetal,  
Universidad de Murcia,  
30071 Murcia, Spain*

**SIR** — While the Convention on Biological Diversity has the laudable goal of increasing the income of the poor countries, we have seen no public discussion in *Nature* or elsewhere of possible adverse consequences, particularly in countries where governments have ownership rights over plant (or other) species.

Payments by foreigners directly to governments or countries in which plant species are found could have a significant antidemocratic effect by providing a source of income independent of control by a country's citizens. In a democratic country, one important way in which citizens can control the government is through action to limit the amounts and use of governmental income that is obtained, for example, from taxes. Once governments have access to a significant source of income independent of its citizens, it reduces the control citizens may exert through their restriction of government access to money. A government could more easily take unpopular action because it would be less important whether citizens would willingly support such action.

It may be argued that 50 per cent (or some other percentage) of such monies will have to be applied to conservation. But what guarantees are there that countries will adhere to such restrictions? And what can be done if they don't? Invade them? Even if followed rigorously, this restriction still leaves a significant amount of discretionary income to governments. If these kinds of 'royalty' payments for use of plant species are to be made at all, we think it would be much safer, from a democratic point of view, to make them directly to a country's inhabitants, not to their rulers.

A good example of what might happen is what actually did happen in the case of the Organization of Petroleum Exporting Countries (OPEC). A country that 'owned' certain valuable plant species as a monopoly (or, in the case of dispute co-owned them along with a few other countries) could act as OPEC

does. In OPEC countries, the politicians are far more eager to please the politicians of other OPEC countries than they are to please their own citizens, because the government's income depends to a far larger extent on the cooperation of the rulers of other OPEC countries than it does on that of their own citizens. Do we really want to see this model of governance spread to other countries? Developing countries are greatly in need of more democracy, not less.

**Durk Pearson**

**Sandy Shaw**

*PO Box 2160,  
Tonopah, Nevada 89049, USA*

# Missing the Moon

**SIR** — Your leading article (*Nature* 358, 609; 1992) criticizing the obsession of the National Aeronautics and Space Administration (NASA) with an enormously expensive but unproductive manned space programme — while solemnly insisting that this obsession is a genuine mistake on NASA's part and that "everybody's interests, NASA's included, would be better served" if it changed its ways — displays a rather touching innocence. The truth, as George Keyworth pointed out during his tenure as President Ronald Reagan's science adviser, is that NASA deliberately misleads Congress, the White House and the public far more than any other US government agency — and it does so simply because its most important real interest by far (like that of any other government agency) is to maintain its current size and employment level, while it has far less real justification for its continued existence at such current levels (left over from the days of the Moon race) than any other US government agency. Naturally, it will sensibly continue pulling the wool over congressional and executive eyes for as long as it can get away with it, which is likely to be for a very long time.

One way to improve the situation might be to strip NASA of all its ability to select space science projects (manned and unmanned) by transferring responsibility to other agencies such as the National Science Foundation, the National Institutes of Health, the National Oceanographic and Atmospheric Administration and the Departments of Commerce and the Interior — but any real major improvement will come only when Congress and the White House contain a large number of people with a reasonable level of basic scientific knowledge, which I expect to see the week of the millennium.

**Bruce Moomaw**

*2953 Oakleaf Drive,  
Cameron Park, California 95632, USA*



## The brain size/IQ debate

**SIR** — Some of your correspondents<sup>1–3</sup> minimize observed race differences in brain size. But such differences are often observed even before statistical correction and the IQ–brain size relationship is robust even among children.

In my US military study<sup>4</sup>, Asian Americans had larger *uncorrected* cranial capacities than African Americans across all male/female, officer/enlisted categories despite being shorter in stature and lighter in weight. Racial differences show up early in life. In a US national study, 17,000 white infants and 7-year-olds had significantly larger head perimeters than their 19,000 black counterparts even though, by 7 years, black children were taller and heavier<sup>5</sup>. In all groups, head perimeter at birth and at age 7 correlated with IQ at age 7 from 0.10 to 0.20. Small differences in brain volume can translate into millions of excess neurons<sup>6</sup>.

Brand<sup>3</sup> claimed that nothing of any theoretical significance has been learned. But brain size has become the main constraint in primate life history theory<sup>7,8</sup>. I have proposed that these theories explain measured differences among human populations<sup>9</sup>.

**J. Philippe Rushton**

*University of Western Ontario,  
Department of Psychology,  
Social Science Centre,  
London, Ontario,  
Canada N6A 5C2*

**SIR** — It is indisputable that men's brains are, on average, larger than those of women. Data from Ho *et al.*<sup>10</sup> show that at the age of 25, when human brains are at maximum size, brains of men weigh about 175g (17 per cent) more than those of women. This large sex-difference in brain mass has long been attributed to the larger body size of men. My analyses<sup>11</sup> have shown (as have, ironically, those of Schluter<sup>1</sup>) that this explanation is false. Statistically, less than 15 per cent of variation in brain mass among men and among women is explained by variation in body size. Clearly, most brain-size variation in humans is due to something other than body size.

Surely, no one would claim that the sex-difference in brain size is due to, for example, girls having poorer nutrition than do boys in North America or Western Europe. Therefore, I propose that the difference is genetically based and probably mediated by the male sex-hormone testosterone (which is responsible for many other sex differences). Only further research can refute my hypothesis that the sex-difference in brain size is related to men's, on aver-

age, greater spatial and mathematical reasoning abilities.

**C. Davison Ankney**

*Department of Zoology,  
University of Western Ontario,  
London, Ontario N6A 5B7, Canada*

1. Schluter, D. *Nature* **359**, 181 (1992).
2. Blest, A. D. *Nature* **359**, 182 (1992).
3. Brand, C. R. *Nature* **359**, 768 (1992).
4. Rushton, J. P. *Intelligence* **16**, 401–413 (1992).
5. Broman, S. H., Nichols, P. L., Shaughnessy, P. & Kennedy, W. *Retardation in Young Children* (Erlbaum, Hillsdale, NJ, 1987).
6. Haug, H. *Am. J. Anat.* **180**, 126–142 (1987).
7. Harvey, P. H. & Krebs, J. R. *Science* **249**, 140–145 (1990).
8. Dunbar, R. I. M. *J. hum. Evol.* **20**, 469–493 (1992).
9. Rushton, J. P. *Can. Psychol.* **32**, 29–42 (1991).
10. Ho, K., Roessmann, U., Straumfjord, J. V. & Monroe, G. *Arch. Pathol. Lab. Med.* **104**, 635–640 (1980).
11. Ankney, C. D. *Intelligence* **16**, 329 (1992).

□ This correspondence is now closed.  
— Editor, *Nature*.

## Human rights

**SIR** — The author of the News article “Chinese scientist arrested for entrepreneurial spirit” (*Nature* **359**, 763; 1992) suggests that “clearly, what is needed is an ongoing educational campaign about the value of knowledge”. What is lacking in China is a proper judicial system where human rights are respected. In a totalitarian political system where the law courts are just the tools of the ruling Communist party, one cannot expect the judiciary to respect the rights of any individual, and certainly not scientists. Clearly what is needed is a campaign to educate the whole country, especially the government, about the idea of law, and the institution of an effective legal system independent of any political party.

**P.-L. Chau**

*University of Cambridge,  
Department of Pharmacology,  
Cambridge CB2 1QJ, UK*

**SIR** — As a Chinese student, I wish to point out that the “relaxed rules for visiting citizens” by the Chinese authorities (see *Nature* **358**, 98; 1992) do not apply to Chinese citizens abroad. In particular, these rules do not apply to those once involved in the democratic movements of Chinese students and scholars. Thus a Chinese student at Harvard University, Ms Gong Xiaoxia, was expelled by the police when she arrived in Guangzhou to visit her family on 28 May. Moreover, the rules do not apply to Chinese students and scholars sent by the Chinese government but to visiting citizens who hold ‘private passports’.

**H. Yang**

*University of California, Los Angeles,  
Los Angeles, California 90024, USA*

## Climate change

**SIR** — The Earth Summit in Rio de Janeiro and the earlier struggle for a Convention on Climate Change may serve as a reminder that the 1982 Convention on the Law of the Sea has its tenth anniversary on 10 December. It is not only one of the most comprehensive and strongest international treaties ever negotiated but the best possible legal means to protect the global climate. But sadly, there has been little interest in using it for this purpose. For too long, climate has been defined as the average weather and Rio was not able to define it at all. Instead, the Climate Change Convention uses the term ‘climate system’, defining it as “the totality of the atmosphere, hydrosphere, biosphere and geosphere and their interactions”. All that this boils down to is ‘the interactions of the natural system’. What is the point of a legal term if it explains nothing?

For decades, the real question has been who is responsible for the climate. Climate should have been defined as ‘the continuation of the oceans by other means’. Thus, the 1982 Convention could long since have been used to protect the climate. After all, it is the most powerful tool with which to force politicians and the community of states into actions.

**Arnd Bernaerts**

*Jungfernstieg 41,  
2000 Hamburg 36,  
Germany*

## Watch that dog!

**SIR** — Diane Gershon (*Nature* **358**, 614; 1992) discusses the allegation that use of recombinant bovine growth hormone (rbGH) would result in increased use of antibiotics to treat bovine mastitis “and therefore higher antibiotic residues in meat and milk”. This is unlikely. Cows being treated for mastitis are not used for beef, and their milk cannot be marketed during the treatment period or for several days afterwards. The matter was resolved many years before the advent of rbGH.

In this and another recent story (*Nature* **358**, 4; 1992), Jeremy Rifkin's group is described as a “consumer watchdog”. Rifkin's various anti-technology campaigns should be viewed in the light of his avowed objective, repeatedly stated in his books, that the world population should be reduced to one billion and that work should revert to human and animal labour.

**Thomas H. Jukes**

*University of California, Berkeley,  
6701 San Pablo Avenue,  
Oakland,  
California 94608, USA*

# Mir for the crystallographers' money

Barry L. Stoddard, Roland K. Strong, Anthony Arrott & Gregory K. Farber

**It is nearly 10 years since the first attempts were made to grow better protein crystals in the low gravity environment of space. Can the enormous cost of these experiments be justified by the results?**

THE idea of growing crystals in a microgravity environment remains a controversial topic. At the heart of the controversy are suggestions that such experiments are primarily funded so that government space agencies can claim to be supporting cutting-edge research in either biotechnology (protein-crystal growth) or in materials science (defect-free silicon crystals, for example). This can then be used to justify expensive new facilities (such as the US space station *Freedom*). What is the status of protein-crystal growth in microgravity? The evidence from such experiments unambiguously shows that growth in microgravity does have the possibility to alter crystal size, quality and morphology. However, experiments to date have not yet demonstrated that microgravity protein-crystal growth is a wise way to spend ever more scarce government research dollars.

The first protein-crystal growth experiment was conducted aboard Space Lab 1 in 1983 (refs 1,2). This experiment illustrates the dangers inherent in reporting microgravity protein-crystal growth experiments. The authors of this study grew lysozyme crystals 1,000 times larger in volume than crystals grown in the same device on Earth, and their report is still cited as evidence that microgravity can increase crystal size. Unfortunately, the authors failed to point out that the crystals they grew were no larger than the average lysozyme crystal grown using standard techniques in an Earth laboratory. What they demonstrated was that their specialized crystal-growth device intrinsically limited the size of lysozyme crystals on Earth. Microgravity overcame some of these limitations, but the overall effect was no improvement in lysozyme crystal size. Clearly, to get an unbiased view, crystals grown in space, crystals grown on Earth in the microgravity hardware, and the best Earth-grown crystals that come from typical laboratory devices must be compared.

## Experimental problems

These control experiments are costly in terms of protein and of time and effort. For crystallization experiments that take advantage of crystal growth geometries available only in microgravity, these control experiments are not possible. For these reasons, not all microgravity experiments have reported

proper controls. In addition to the difficulty with control experiments, growing crystals in microgravity is very different from growing them on Earth. The apparatus must be compatible with a microgravity platform. Transport of proteins and apparatus to the launch site and setting-up in the vehicle must be carefully planned, which is especially problematic when launch delays occur. Mistakes in initiating or monitoring an experiment in microgravity will render the experiment useless. These problems, combined with the great expense of an experiment, have raised questions in the crystallographic community about the wisdom of performing such experiments.

Despite these questions, several groups have reported experiments on different microgravity platforms. An extensive series of experiments have been flown on the US Shuttle<sup>3-8</sup>, one experiment has been performed on a Chinese *Long March* rocket<sup>9</sup>, and we have flown two long-duration experiments on the Russian space station *Mir*<sup>10,11</sup>. Although the Shuttle experiments are continuing and complete results from these experiments have not yet been published, the results from recent flights have been summarized<sup>5,7</sup>. Averaged over four missions, the Shuttle experiments show that 20% of crystals are an improvement on their Earth-grown counterparts, 40% produced crystals too small for data collection or showed no improvement, and the remaining 40% produced no crystals at all<sup>5</sup>. Crystals which improved in the Shuttle experiments include interferon- $\gamma$  (larger crystals, higher scattering power), elastase (larger crystals, higher scattering power), isocitrate lyase (change from a dendritic to a prismatic morphology), *Lathyrus ochrus* lectin I (higher resolution data), canavalin (more uniform morphology, higher scattering power), phospholipase A2 (more uniform morphology) and serum albumin (thicker plates). Typical improvements were an increase in resolution of 0.2–0.3 Å and an increase in the number of observable reflections of 10–15%. These numbers are comparable to the results from our second *Mir* experiment (see table overleaf).

The complete results from our first *Mir* experiment have been published<sup>11</sup>. Preliminary results of our most recent experiment are summarized in the table: 5 of the 21 proteins in this experiment

produced crystals that were clearly superior to their Earth-grown counterparts and in which the improvement could be attributed only to microgravity (24%). Microgravity had no effect on six other crystallization experiments (29%), and for seven of the proteins the effects of microgravity were unclear (33%). Some of the reasons for these ambiguous results were poor experimental design and incomplete ground control experiments. However, in several cases, conditions which usually produced crystals on Earth did not produce crystals in microgravity. Crystals of one protein were clearly worse (5%), and the results for two proteins are still unknown (10%). Overall, our results are similar to the Shuttle experiments both in the frequency and type of improvements we observe.

## Improved crystals

Three different types of improvements are seen in both the Shuttle and *Mir* experiments: larger crystals, crystals with improved scattering power, and new crystal morphologies. The improvements in size and quality have been attributed, at least in part, to the lack of convective mixing in microgravity. During crystal growth, the solution immediately adjacent to the growing crystal face is depleted of protein. This solution is less dense than the bulk solution, so it rises in a gravity field. These convection currents can be dramatic<sup>12,13</sup>, but the effect of these currents on crystal growth and quality could not be assessed until experiments could be conducted in the absence of gravity. The effect of microgravity on crystal morphology is just as important as the changes in crystal size and scattering power. So far, very different morphologies have been observed for lysozyme<sup>11</sup>, 10-tetra-hydrofolate synthetase and isocitrate lyase<sup>4</sup>.

We conclude from these experiments that microgravity acts in much the same manner as any of the standard chemical additives in crystallization experiments. That is to say that it often has no effect, but when it does have an effect, it can be either an improvement or a lessening in the quality of the obtained crystals. Microgravity is not a universal panacea for crystals unsuitable for diffraction experiments. But considering the limited number of experiments performed so far and the difficulties inherent in performing them, the observed improvement



## EXPERIMENTS WITH DIFFRACTION-DATA QUALITY CRYSTALS AVAILABLE IN LAB

| Protein                                                          | Average lab crystals                                   | Microgravity crystals                          | Effect |
|------------------------------------------------------------------|--------------------------------------------------------|------------------------------------------------|--------|
| 1. Anti-Brucella Fab (Monoclonal)                                | 0.6 × 0.3 × 0.3 mm<br>$d_{\min} = 3.7 \text{ \AA}$     | Larger<br>Higher resolution (0.15 Å)           | +      |
| 2. Fab – HPr complex (Monoclonal)                                | 0.5 × 0.25 × 0.15 mm<br>$d_{\min} = 2.8 \text{ \AA}$   | Same size, single<br>Higher resolution (0.1 Å) | +      |
| 3. Isocitrate dehydrogenase ( <i>Escherichia coli</i> )          | 0.6 × 0.6 × 0.3 mm<br>$d_{\min} = 2.5 \text{ \AA}$     | Smaller<br>Higher resolution (0.2 Å)           | +      |
| 4. $\beta$ -lactamase ( <i>Enterobacter cloacae</i> )            | 1.0 × 0.15 × 0.01 mm<br>$d_{\min} = 2.3 \text{ \AA}$   | Smaller                                        | 0      |
| 5. $\beta$ -lactamase ( <i>Bacillus licheniformis</i> )          | 0.75 × 0.25 × 0.005 mm<br>$d_{\min} = 2.0 \text{ \AA}$ | Smaller                                        | 0      |
| 6. Lysozyme (Hen egg white)                                      | 1.0 × 0.5 × 0.5 mm<br>$d_{\min} = 1.9 \text{ \AA}$     | Same size<br>$d_{\min} = 1.9 \text{ \AA}$      | 0      |
| 7. Isocitrate dehyd. (complex) ( <i>E. coli</i> )                | 0.5 × 0.5 × 0.2 mm<br>$d_{\min} = 2.8 \text{ \AA}$     | None obtained                                  | ?      |
| 8. Trp tRNA synthetase ( <i>E. coli</i> ) (Apo and tRNA complex) | 0.3 × 0.1 × 0.1 mm<br>$d_{\min} = 3.5 \text{ \AA}$     | None obtained                                  | ?      |
| 9. Anti-chymotrypsin Human                                       | 0.3 × 0.3 × 0.1 mm<br>$d_{\min} = 3.5 \text{ \AA}$     | None obtained                                  | ?      |
| 10. D-Ala-D-Ala ligase ( <i>E. coli</i> )                        | 0.4 × 0.15 × 0.03 mm<br>$d_{\min} = 2.1 \text{ \AA}$   | Smaller<br>$d_{\min} = 4 \text{ \AA}$          | ?      |
| 11. D-Ala-D-Ala ligase ( <i>Salmonella typhimurium</i> )         | 0.5 × 0.2 × 0.2 mm<br>$d_{\min} = 2.5 \text{ \AA}$     | None obtained                                  | ?      |
| 12. $\beta$ -hexosaminidase                                      | 0.3 × 0.1 × 0.1 mm<br>$d_{\min} = 3.2 \text{ \AA}$     | None obtained                                  | ?      |
| 13. Non-fluorescent flavoprotein                                 | 0.3 × 0.12 × 0.03 mm                                   | Smaller                                        | –      |

## EXPERIMENTS WITHOUT DIFFRACTION-DATA QUALITY CRYSTALS AVAILABLE IN LAB

| Protein                                      | Average lab crystals                                 | Microgravity crystals                                 | Effect |
|----------------------------------------------|------------------------------------------------------|-------------------------------------------------------|--------|
| 1. Aspartase ( <i>E. coli</i> )              | 0.2 × 0.02 × 0.02 mm<br>$d_{\min} = 4.0 \text{ \AA}$ | Larger<br>Higher resolution (0.8 Å)                   | +      |
| 2. Tetrahydrofolate synthetase (Clostridium) | 0.3 × 0.1 × 0.02 mm<br>No measurable diffraction     | Larger, new morphology<br>$d_{\min} = 30 \text{ \AA}$ | +      |
| 3. Aspartate receptor ( <i>E. coli</i> )     | None currently obtained                              | None obtained                                         | 0      |
| 4. Luciferase (Firefly)                      | 0.5 × 0.005 × 0.005 mm<br>No measurable diffraction  | None obtained                                         | 0      |
| 5. Histidase                                 | None currently obtained                              | None obtained                                         | 0      |
| 6. Rhodopsin (Bovine)                        | 0.05 × 0.05 × 0.05 mm<br>No measurable diffraction   | Smaller                                               | ?      |

Sources of protein: George Birnbaum, National Research Council of Canada; David Christianson, Univ. Pennsylvania; Wim deGrip, Univ. Nijmegen; Louis Delbaere, Univ. Saskatchewan; G. K. F.; Michael N. G. James, Univ. Alberta; James Knox, Univ. Connecticut; Daniel Koshland, Univ. California, Berkeley; Allen Phillips, Pennsylvania State Univ.; Jesse Rabinowitz, Univ. California, Berkeley; Ronald Viola, Univ. Akron; Daniel Yang, McMaster Univ. In addition to the proteins listed above, two proteins were flown by a company which prefers not to reveal the name of the proteins. One of them is an integral membrane protein, the other is soluble. In neither case were the results of the experiment revealed to us.

## Effects:

+, Measurable improvement in size or diffraction strength or change in morphology attributed solely to the effect of microgravity as an additive in the crystallization.  
–, Measurable deterioration in crystallization of macromolecule attributed solely to the effects of microgravity.  
0, No discernible effect of microgravity on crystallization.  
?, Effect of microgravity not determinable through the results obtained.

rate of 20% probably underestimates the number of crystallization problems that would benefit.

Microgravity has not yet accomplished any significant breakthrough in protein-crystal growth. So far, no protein has been reported to crystallize in microgravity that does not crystallize on Earth. In addition, although several protein crystals have been reported to show dramatic improvements in a microgravity environment ( $\beta$ -galactosidase<sup>1</sup>, isocitrate lyase<sup>4</sup> and aspartase, for example) none of these crystal structures has been solved. Microgravity-grown crystals were used in determining the structure of interferon- $\gamma$ , but the microgravity crystals were not essential to the solution of the structure<sup>14</sup>.

The next steps are obvious. Because

we cannot predict if a given crystallization problem will fall into the 20% that improve in microgravity, we need to develop criteria to aid in this determination. The effect of microgravity on proteins which do not crystallize on Earth should be explored. Experiments should be done to determine how often microgravity is able to turn small crystals which diffract to low resolution into crystals which diffract strongly enough to solve structures. We plan to address these problems on our next flights on *Mir*. We hope that readers with proteins which may be appropriate for such experiments will contact us.

Measured by the yardstick of routine crystal production, microgravity is not yet a success. Given the difficulty and cost of such experiments, we doubt that

it will be appropriate for many protein-crystallization problems. The final basis by which to assess its importance and usefulness will be the ability to solve protein structures using crystals which could only be grown in microgravity.

Based on our 24% success rate, the experiments reported here show that for protein-crystal growth, *Mir* is as good a microgravity platform as the Shuttle. Because the Russian launch vehicle which delivers our experiments to *Mir* is unmanned, we have found that sending experiments to *Mir* is more reliable than trying to do similar experiments on the Shuttle. Finally, because *Mir* is a permanently orbiting space station, we can do long-duration experiments that are not possible on any other microgravity platform. It is clear that researchers in the West should consider using *Mir* for their microgravity experiments. Given the uncertainties over continued funding for the *Mir* programme in Russia, such collaboration would clearly be of mutual benefit for crystallographers. □

Barry L. Stoddard is at the Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104; Roland K. Strong is in the Division of Biology, Mail Stop 156–29, California Institute of Technology, Pasadena, California 91125; Anthony Arrott is at Payload Systems, 276 Third Street, Cambridge, Massachusetts 02139; Gregory K. Farber is in the Departments of Chemistry and Molecular and Cell Biology, 152 Davey Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

1. Litke, W. & John, C. *J. Cryst. Growth* **76**, 663–672 (1986).
2. Litke, W. & John, C. *Science* **225**, 203–204 (1984).
3. DeLucas, L. J. *et al. J. Cryst. Growth* **76**, 681–693 (1986).
4. DeLucas, L. J. *et al. Science* **246**, 651–654 (1989).
5. DeLucas, L. J. & Bugg, C. E. *Adv. Space Biol. Med.* **1**, 249–278 (1991).
6. Sibille, L. & Baird, J. K. *J. Cryst. Growth* **110**, 72–79 (1991).
7. DeLucas, *et al. J. Cryst. Growth* **110**, 302–311 (1991).
8. DeLucas, L. J. & Bugg, C. E. *Methods* **1**, 105–109 (1990).
9. Plass-Link, A. & Cornier, J. *J. appl. Microgravity Tech.* **3**, 123–132 (1988).
10. Stoddard, B., Strong, R. K., Farber, G. K., Arrott, A. & Petsko, G. A. *J. Cryst. Growth* **110**, 312–316 (1991).
11. Strong, R. K., Stoddard, B. L., Arrott, A. & Farber, G. K. *J. Cryst. Growth* **119**, 200–214 (1992).
12. Baird, J. K., Meenan, E. J., Xidis, A. L. & Howard, S. B. *J. Cryst. Growth* **76**, 694–700 (1986).
13. Shlichta, P. J. *J. Cryst. Growth* **76**, 656–662 (1986).
14. Eallick, S. E. *et al. Science* **252**, 698–702.

ACKNOWLEDGEMENTS. We thank the technical staff of Payload Systems, Inc. for support, including Maria Douglass, Christopher Krebs, Julianne Zimmerman and Bob Renshaw. Soojay Banerjee, Beth Buelow, and Richard Kidd helped with crystal and data collection at Penn State. Some crystallographic data collection was performed with the assistance of Bob Sweet and Paul Singer at the National Synchrotron Light Source, and in the lab. of Sung-Hou Kim at the University of California, Berkeley. We thank the technical staff of NPO-Energia, especially Olyeg Mitichkin. Operations on *Mir* were performed by cosmonauts Alexander Volkov and Sergei Krikalov. G.K.F. is supported by the Searle Scholar Foundation and the PRF. During these experiments, B.L.S. was supported in the lab of Daniel Koshland by the Helen Hay Whitney Biomedical Research Foundation. R. K. S. is supported in the lab of Pamela Bjorkman by the American Cancer Society. We also acknowledge Mary Chung. Address correspondence to G.K.F.

# Molecular neurology comes of age

**Participants at *Nature's* conference on "The brain in well-being and disease" (12–13 November) were excited that hitherto intractable problems are now beginning to crack under the impact of powerful new techniques.**

**Boston.** Four years ago, at *Nature's* last neuroscience conference, Maxwell Cowan (Howard Hughes Medical Institute) presciently remarked that instead of exploring information processing in the brain with models such as neural networks, *Nature's* next neuroscience conference would centre on questions that could be answered only with the aid of molecular biology. He was right. Of course, predictions that a given approach will transform a field are nothing new, especially in neurobiology. But as Colin Blakemore (University of Oxford) observed in one of his session summaries (intended to raise the questions "no-one else would ask"), the nervous system was produced by evolution, the basis of which is genetic, so that the capacity of molecular genetics to inform neurobiology must be vast. This year's conference even produced tantalizing hints that these techniques could eventually revolutionize treatment of neurological and mental disorders, which currently afflict more than 10 per cent of Western populations.

Of course, by no means every talk was dominated by molecular biology. Marcus Raichle (Washington University) focused on a top-down approach, using non-invasive techniques to image whole, functioning brains. By summing repeated differential PET scans of subjects performing tasks such as recognizing words (real or nonsense), the induced changes in brain metabolic activity can be pinpointed. With repetition, fewer regions of the brain are involved, and the response becomes almost reflex, suggesting that absent-minded slips need not be Freudian.

The importance of an integrated approach was also apparent in the account by David Hubel (Harvard Medical School) of how a monkey perceives stereoptic depth. The mathematical basis for this was discovered by Wheatstone in 1838 (although Leonardo da Vinci only narrowly missed it), but the video with which the talk closed (already evidently a classic featuring in at least one undergraduate course) showed elegantly how the theory is realized in neurons responding only to lines at an appropriate angle and distance.

In other talks about the senses, the molecular bias was inescapable. Indeed, one of the most fascinating themes to emerge was that hearing, sight and smell use strikingly different molecular mechanisms to generate internal representations of external events. In the ear, cellular responses three orders of magnitude faster than nor-

mal are necessary to detect sounds at 20 kilohertz, precluding the use of a second messenger system. As James Hudspeth (University of Texas Southwestern Medical Center, Dallas) explained, the solution is to couple hair-bundle deflection to membrane ion channel gating mechanically. In the organ of Corti, ion channels at the tips of short rigid 'hairs' are each connected to the adjacent hair by a thin fibre link. These hairs are grouped in tightly packed bundles and pivot at their bases when deflected by vibrations, so that they slide against one another, tugging on the linking fibres and opening the channels to admit  $K^+$  and  $Ca^{2+}$ . Direct measurements of the hair bundles' resistance to deflection confirm that it is greatest at the point where the channels open, and studies of how this point can be adjusted to control sensitivity (apparently by using a myosin molecule to move the channel itself) are now under way.

Research on vision is beginning to address the distortion of colour perception in certain genetic disorders. The genes encoding the human cone cell receptors for red and green light are very similar, making them prone to unequal recombination (Jeremy Nathans, Johns Hopkins University). Sometimes this simply shuffles exons between the genes, producing receptors with intermediate spectral sensitivities (and people who see colours slightly differently). But on other occasions a 600 base-pair sequence upstream of the genes is deleted, abolishing expression of both genes and causing blue-cone monochromacy. The deleted sequence is now known to contain the locus control region for these genes.

In the visual system, numerous colours are recognized by only three receptor proteins with the aid of lavish signal processing; in the olfactory system, as Blakemore remarked, the complexity resides in the receptors themselves. Olfactory epithelia contain as many as 1,000 different receptor proteins, which constitute a new group within the G-protein coupled, seven-transmembrane-helix family, and otherwise occur only in sperm (Richard Axel, Columbia University). Unlike the immunoglobulins, the different receptors are not apparently produced by genomic rearrangement, but each neuron in the olfactory epithelium may nevertheless express only one or a few receptor genes.

Elsewhere in the brain, the molecular events may be less clearly understood, but molecular tools can be just as valuable. In exploring the development of the nervous

system, the first task is clearly to determine how the necessary numbers and types of cells are produced. A detailed understanding of this process will probably depend on the use of genetically engineered mice, as Mario Cappechi (University of Utah) showed (and scientists should not shrink from publicizing the importance of such work, added Blakemore). Abolishing expression of the homeobox gene *Hox 1.6*, for example, produces animals lacking the fifth hindbrain rhombomere. By injecting fluorescent dyes into these animals' neurons, it is possible to show that the seventh and lower ganglia are asymmetrically shifted anteriorly, altering the innervation of the corresponding targets. Lack of this rhombomere also induces alterations in its higher neighbours.

Using different molecular tools to tackle a related problem, Connie Cepko (Harvard Medical School) has marked embryonic chick retinal cells with retroviruses and transplanted them into embryos of various ages. As development proceeds, progressive changes in progenitor cell behaviour are coupled with alterations in the cues eliciting differentiation (including gross factors and small molecules like taurine) to produce a correctly structured retina.

To reduce the three-dimensional problem of how cell identity is specified to essentially one dimension, Thomas Jessell (Columbia University) has turned to the early stages of spinal-cord development. Differentiation along the dorsoventral axis of the chick neural tube is regulated by two ventral midline cell groups, the notochord and the floorplate. Once floorplate formation has been induced by the notochord, the floorplate is the source of factors responsible both for maintaining its own differentiated state and for inducing cells lateral to it to become motor neurons. This action is opposed by a member of the *TGF $\beta$*  family called dorsalin, which is exclusively produced by cells in the dorsal part of the neural tube, and induces formation of neural crest-like cells.

Once the requisite cell numbers and types are present, how do they link up? To answer this question, Corey Goodman (University of California at Berkeley) has studied motor neurons in developing *Drosophila* segments. Normally, axons grow out towards one of 30 possible target muscles until they are near enough to form synapses. Mutations that affect this process have been found in as many as 15 genes on chromosome 2 alone, suggesting that



the whole genome may yield more than 40. The mutations in question each affect all neurons innervating one or more of the four segmental muscle groups, and fall into three classes: those that arrest axonal growth before the target is reached, those that block the transition from target finding to synapse formation, and (perhaps most interestingly) those that induce axonal sprouting as a result of the neuron's inability to recognize its target.

On reaching its target, the neuron must next form synapses in the correct position. Martha Constantine Paton (Yale University) has studied this process in the developing visual pathways of tadpoles and rats. Although numerous connections are initially formed, the synapses that are simultaneously used activate the NMDA receptor and are rapidly strengthened, while synapses with poorly correlated activity are pruned. The effective input remaining may then be stabilized by downregulating NMDA receptor function, preventing the subsequent development of poorly correlated inputs but decreasing brain plasticity. There is increasing evidence that specific NMDA receptor subtypes play important roles in this process (prompting Blakemore to ask whether we are demanding too much of one receptor).

As formation of these pathways in mammals occurs before they open their eyes, it must rely on the spontaneous activity of retinal ganglion cells. As Carla Schatz (University of California at Berkeley) explained, such activity is essential for the correct patterning of the retinal connections to the lateral geniculate nucleus in cats and ferrets. Although the neurons in this region initially respond to input from both eyes, they later become organized into a series of layers, each accepting signals from one eye alone. How could such a pattern develop? Multichannel recordings of the electrical activity in isolated retinal fragments show that random waves of activity spread across each retina every few minutes at speeds of up to 300  $\mu\text{m}$  per second. Such activity means that input from both eyes virtually never reaches a target neuron simultaneously, supporting the hypothesis that synapses asynchronous with each neuron's strongest inputs are progressively pruned.

It has long been known that more neurons than necessary are generated in development. The unwanted cells die by activating a pathway termed apoptosis, in which the cell fragments its DNA, shrinks, and is rapidly phagocytosed by neighbouring cells. It has only recently become clear, however, that such deaths are very common, and are by no means confined to neurons. By implanting transfected Cos cells secreting PDGF into the rat optic nerve, Martin Raff (University College, London) has succeeded in preventing more than 80 to 90 per cent of the oligodendrocyte deaths that would otherwise occur. As well as confirming that these deaths are a direct

result of trophic-factor deprivation, this makes it possible to estimate that as many as half the oligodendrocytes formed in the optic nerve eventually die.

How is this process controlled? As Robert Horvitz (Massachusetts Institute of Technology) demonstrated, the nematode *Caenorhabditis elegans* provides an excellent system with which to address this question, since all 959 cells in the adult (as well as the 131 that die during development) are of known lineage. In this organism, cell death is initiated by the products of the genes *ced3* and *ced4*, following which the products of additional genes mediate phagocytosis and DNA degradation. If the product of either gene is inactive in a cell, that cell will survive even when transplanted into a wild-type host. The functions of these proteins are unknown, although the *ced3* product may well be a target for phosphorylation, and the *ced4* product contains two calcium-binding motifs. A further gene involved in this process, *ced9*, has the opposite effect, as without its product many or all cells perish at an early stage. The sequence of the *ced9* protein is 23 per cent identical to that of the mammalian oncoprotein bcl2, which is also known to play a crucial role in preventing apoptosis, and remarkably can substitute for *ced9* in the worm. The precise function of both proteins, however, remains unknown.

In searching for the molecular basis of apoptosis, Eugene Johnson (Washington University School of Medicine) has shown that the survival of neurons deprived of nerve growth factor increases as the intracellular calcium concentration is raised. So far, no genes have been directly implicated in the death of these cells, although it is intriguing that irreversible commitment to apoptosis is marked by an increase in the level of cyclin D<sub>1</sub> (usually associated with the G<sub>1</sub>- to S-phase transition in the cell cycle).

If the techniques of molecular biology have transformed neurobiological research, will they have a comparable impact on clinical practice? The indications are that they will. The possibility that inhibitors of nitric oxide production may one day transform the treatment of stroke is proof enough of that. As Solomon Snyder (Johns Hopkins University) recounted, nitric oxide was originally identified as endothelial-derived relaxing factor, a substance vital for the control of blood pressure, and the discovery that it is also a neurotransmitter proved that neurotransmitters need not be stored in vesicles, released by exocytosis into the synaptic cleft, or taken up by specific membrane-bound receptors. The clinical relevance of this work, however, is that inhibitors of nitric oxide production strongly inhibit cell death after simulated strokes, implying that such death can result from nitric oxide release by ischaemic neurons ("Is this related to apoptosis?", asked

Blakemore). The hunt is now on for inhibitors of the neuronal nitric oxide synthase that do not affect the endothelial enzyme.

Another scourge of our ageing population is Alzheimer's disease, and here too there has been exciting progress. Although it is clear that the 40-amino-acid peptide cleaved from the  $\beta$ -amyloid precursor protein is central to the pathogenesis of the disease, the route by which it is generated remains unclear. The recent finding that cultured cells transfected with the gene encoding the precursor protein secrete small quantities of the peptide (Dennis Selkoe, Harvard Medical School) is therefore particularly welcome, as it should facilitate a search for inhibitors of peptide production. A mutation in the precursor protein found in a Swedish case of familial Alzheimer's disease has already been shown to cause increased production of peptide by transfected cells.

In other cases, the basis of a disease becomes clear on isolation of the relevant gene, as happened with the disorders that show 'anticipation' (becoming more severe in successive generations). In myotonic dystrophy, it is now clear that this phenomenon results from the CTG repeats in the 3' untranslated region of the affected gene (David Housman, Massachusetts Institute of Technology). The commonest allele has only five such repeats, and is stable, but alleles with sixty repeats are not uncommon, and occasionally undergo slippage during replication. The consequence is an increasing number of repeats in each affected generation, leading to progressive worsening of the phenotype. A similar process occurs in the commonest inherited cause of mental retardation, fragile X syndrome.

Can the techniques of molecular biology also suggest solutions to previously intractable clinical problems, such as the inability of the central nervous system to regenerate following injury? Fred Gage (University of California at San Diego) has approached this question by infecting fibroblasts with a disabled retrovirus encoding nerve growth factor and implanting them near the transected axons of cholinergic neurons projecting from the septum to the hippocampus. In this way, cells that would otherwise die can instead be induced to regenerate. Remarkably, as the axons grow through the fibroblast implant, they become sheathed by astrocytes, which appear to be the preferred substrate for axonal growth, rather than inhibiting it as had previously been thought. Of course, the prospect of reversing the consequences of a head or spinal cord injury by a carefully placed injection of cells secreting a suitable cocktail of trophic factors is still remote, and the many advances required to accomplish this goal will doubtless be a fertile source of future *Nature* conferences. Nevertheless, that such a feat is a conceivable at all testifies how dramatically Cowan's prediction has been fulfilled.

Nicholas Short

# Keeping the Sun in proportion

Andrew A. Lacis and Barbara E. Carlson

GLOBAL warming observed over the past century has focused attention on man-made greenhouse gases, presumed to be the likely cause. Some researchers, however, have expressed doubt, wondering whether the Sun — our ultimate source of energy — might actually be variable (as stars similar to the Sun are found to be), and thus account for the bulk of the observed climate change. The expressed hope is that future warming due to further greenhouse-gas emissions might therefore not be significant. Two papers in the issue<sup>1,2</sup> put this idea to the test and find it wanting.

The annually averaged global-mean surface air temperature rose by 0.8 °C between 1880 and 1990. This change is broadly accepted, despite the inherent problems of incomplete global coverage, changing observation practices and urbanization. Increases in greenhouse gas concentrations, CO<sub>2</sub>, CH<sub>4</sub>, NO<sub>2</sub> and chlorofluorocarbons have been accurately measured over the same period. Their opacities, which account for the 2 W m<sup>-2</sup> radiative forcing of the climate incurred since 1850, are also accurately known. Reflecting the accelerating pace of industrialization and population growth, 23 per cent of this climate forcing has been incurred over the past decade alone<sup>3</sup>.

However, the observed decadal and interdecadal variability of the global mean temperature are clear indications that factors other than greenhouse-gas forcing must also have contributed significantly to the observed temperature change. In particular, the sharp rise in mean temperature in the 1920s and the subsequent decline during 1940–70 are not at all consistent with the steady increase in greenhouse-gas levels (see Fig. 2 of Schlesinger and Ramankutty<sup>2</sup> on page 332). This inability to model the temperature record with greenhouse-gas forcing alone has led some investigators to question model results of the greenhouse effect and to suggest solar variability as an alternative explanation.

## Maunder Minimum

Although investigating the past behaviour of the Sun is somewhat speculative, the period from 1645 to 1715, known as the Maunder Minimum, was nevertheless particularly striking, in that all sunspot activity ceased and the world experienced simultaneously a cold period referred to as the Little Ice Age. With no other source of readily identifiable climate forcing to explain this unusual period of cold temperatures, some researchers were led to postulate

that changes in the solar output of the order of 0.5 per cent might provide an explanation. Thus it was suggested<sup>4</sup> that the Sun may go through cycling and noncycling phases with the noncycling phase of the Maunder Minimum corresponding to reduced solar output. Using Ca II H- and K-line emission as an index of the Sun's brightness, Lean *et al.*<sup>5</sup> matched and analysed the emissions observed for noncycling stars, finding that the solar irradiance may have decreased below its current mean level by 0.24 per cent during the Maunder Minimum. The decrease in global mean temperature due to this reduction in total solar irradiance is in the range 0.2–0.6 °C, comparable to the estimated 1 °C decrease associated with the Little Ice Age.

Direct measurement of the total solar irradiance during the past sunspot cycle has revealed variations of the order of only 0.1 per cent. The larger variability of the solar irradiance, operating over timescales of decades and centuries, must necessarily be inferred indirectly by correlating some measure of solar change with the historical temperature record. The more successful studies have shown that changes in the solar umbra-penumbra ratio, sunspot cycle envelope and, more recently, changes in length of the solar-cycle period can be closely correlated with the global temperature trend<sup>6</sup>. Although variations in the umbra-penumbra ratio, sunspot envelope, and length of the solar cycle may serve as a useful proxy for changes in solar magnetic-field strength and convective activity that act to modulate solar irradiance, the lack of a clear physical mechanism is a definite drawback.

Kelly and Wigley have now used an energy-balance model to investigate the effects of greenhouse and solar forcing over the period 1765–1985. Their climate simulations, on page 328 of this issue<sup>1</sup>, provide circumstantial support for a link between solar-cycle length and irradiance, but show the solar contribution to the global-mean temperature change to be much less than that due to increasing greenhouse gases and other anthropogenic effects. Schlesinger and Ramankutty, on page 330 of this issue<sup>2</sup>, use a hemispherically resolved version of their climate-ocean model to investigate the effect of greenhouse gas, aerosol and solar forcing over the period 1638–1985. They also conclude that greenhouse gases will continue to dominate global warming, but note that changes in solar irradiance modify the predictions and, in fact, decrease the predicted temperature

change associated with a doubling of atmospheric CO<sub>2</sub> by nearly a half.

Kelly and Wigley provide the caveat that the solar-irradiance record is sensitive to the filtering method used to construct the record from the cycle-length data, limiting the utility of this approach until the physical mechanism linking the solar-cycle length to solar irradiance is better understood. An additional caveat is that spectral measurements of the solar irradiance changes during the past sunspot cycle show that the solar ultraviolet radiation takes a disproportionate share of the variability. Wavelengths short of 300 nm contribute 1 per cent to the total irradiance, but they contribute 20 per cent of the observed variability<sup>7</sup>.

## Ultraviolet radiation

An increase in solar ultraviolet radiation, which is absorbed primarily in the stratosphere, would cause local stratospheric heating with the excess thermal energy being radiated directly to space. This produces no net forcing at the tropopause, hence no increase in surface temperature. Thus, actual solar irradiance variations may be less effective than model results, which assume spectrally uniform changes, imply.

Solar monitoring is clearly needed to establish the long-term variability of solar ultraviolet irradiance. Two solar spectral irradiance experiments (SUSIM and SOLSTICE) on the Upper Atmosphere Research Satellite (UARS) will measure the ultraviolet irradiance within the critical 115–415 nm spectral range. Along with the active cavity radiometer ACRIM II, also on UARS, there is reason to hope that, for the time being, solar-irradiance will be adequately monitored. But to establish the solar-irradiance-climate connection properly will take continuous monitoring over at least several decades. In the meantime, it should be recognized that of the different contributors to climate change, the greenhouse effect of trace-gas forcing is the most accurately documented and the best understood. Policy makers should not be distracted from their vital task of controlling greenhouse-gas emissions. □

Andrew A. Lacis and Barbara E. Carlson are at the Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA.

1. Kelly, P. M. & Wigley, T. M. L. *Nature* **360**, 328–330 (1992).
2. Schlesinger, M. E. & Ramankutty, N. *Nature* **360**, 330–333 (1992).
3. Hansen, J., Lacis, A. & Prather, M. J. *Geophys. Res.* **94**, 16417–16421 (1989).
4. Baliunas, S. & Jastrow, R. *Nature* **348**, 520–523 (1990).
5. Lean, J., Skumanich, A. & White, O. *Geophys. Res. Lett.* **19**, 1591–1594 (1992).
6. Friis-Christensen, E. & Lassen, K. *Science* **254**, 698–700 (1991).
7. Lean, J. *Science* **244**, 197–200 (1990).



# Frog ponds and ocean iron

Stuart L. Pimm

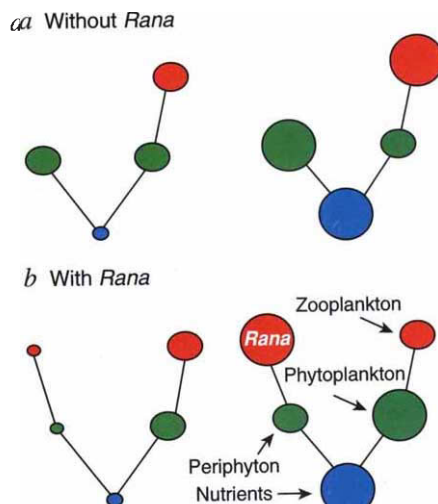
THE way in which perturbations send their effects rippling through ecological systems is exquisitely dependent on the structure and composition of food webs. This is the deceptively straightforward conclusion of Leibold and Wilbur, whose paper (page 341 of this issue<sup>1</sup>) considers the consequence of altering plant productivity.

Suppose we add a limiting nutrient (phosphorus and nitrogen are common choices) to a lake or a field. Perhaps the plants alone will increase in biomass, soaking up atmospheric carbon in the process. If so, should we consider sprinkling iron filings over large areas of the oceans (where phytoplankton are thought to be iron limited<sup>2</sup>) with the hope of ameliorating the increase in atmospheric carbon dioxide? Alternatively, herbivores might consume all the new plant growth, carnivores all the new herbivores, and so on up the food chain. Energetically profligate endotherms sit atop food chains and the carbon could soon be back in the atmosphere. Lotka-Volterra models with one species at each trophic level predict that herbivore biomass will increase only in food chains with two levels (plants, herbivores), and will not increase when there are also carnivores<sup>3</sup>. Studies across natural gradients of productivity show that herbivore biomass increases continuously<sup>4,6</sup>. Of course the models may contain unrealistic terms for interspecific interactions, and the assumption of one species per level may not capture the dynamics of complex food webs. What Leibold and Wilbur now show is that food-web structure alters the response to experimental increases in productivity.

Lotka-Volterra equations model trophic transfers using random 'collisions' between predators and prey. With one species at each trophic level, the top species will always increase in biomass with increasing plant productivity. So plants will increase if there are only plants. Herbivores should increase if there are only plants and herbivores. Plants and carnivores, but not herbivores, increase if there are three levels. The patterns alternate depending on whether the number of levels is odd or even. Oksanen *et al.*<sup>7</sup> suggest that the most barren ecosystems support one functional trophic level, with only sporadic, opportunistic herbivory. More productive systems support herbivores that can regulate plant production. Only in the most productive systems are there predators that regulate the herbivores. Over large gradients in productivity, there should be stepwise increases in

plant, herbivore and carnivore biomass as systems change from one functional trophic level to three.

There are at least two limitations to this theory. First, in ratio-dependent models, transfers between levels depend on the ratio of the densities of the consumers and the consumed; increases in productivity lead to increases in plants and herbivores<sup>8</sup>. Second, with a more



Four food webs showing the relative biomass of nutrients, periphyton, phytoplankton, *Rana* tadpoles and zooplankton in the experimental tanks. The circles represent relative biomasses within each group. Without *Rana* (a), increased nutrient levels increase the ungrazed periphyton, the zooplankton, but not the phytoplankton it grazes. With *Rana* (b), increased nutrients result in increases in all groups except the zooplankton. Low nutrient levels are on the left, high nutrient levels on the right.

complicated food web, the simple model predictions need not hold. Food-web structure may matter.

Compilations of terrestrial herbivore biomasses and plant productivities<sup>4,6</sup> show that productivities range from about 10 g to about 1,000 g carbon m<sup>-2</sup> yr<sup>-1</sup>. Aquatic values span a similar range<sup>6</sup>. The interpretation of these figures is, however, the subject of debate. McNaughton *et al.*<sup>4</sup> argue that herbivore biomass (HB) increases much faster than plant productivity (PP),  $HB \propto PP^{1.52}$ . Moen and Oksanen<sup>5</sup> select data with different criteria, and recommend a model with two slopes:  $HB \propto PP^{2.2}$  for the less productive systems, and  $HB \propto PP^{1.01}$  for the more productive ones. Such a stepwise increase is reminiscent of the theory of Oksanen *et al.*, but there are no ranges of productivity where herbivore biomass is constant. The paucity and quality of the data cloud the

interpretations, however. Some ecosystems are possible outliers (a problem also noted by Cyr and Pace<sup>6</sup>). Why some ecosystems are special in this way is not clear. Even excluding them, the range of biomasses for any given productivity often spans two orders of magnitude. Simply, herbivore biomasses vary substantially.

Leibold and Wilbur show that food-web structure may drive some of this variation. They filled 1,000-litre tanks with water and pine straw as a benthic substrate to produce mesocosms known to mimic natural ponds. An assortment of algae and zooplankton were introduced into the tanks, to half of which were added nitrogen and phosphorus to create high nutrient levels. Other factors in the analysis included the presence or absence of *Daphnia* and *Rana* tadpoles. *Daphnia*, like the other zooplankton present, graze the phytoplankton. *Rana* graze the periphyton (the attached algae and associated organisms). The phytoplankton and the periphyton compete for nutrients.

Without *Rana*, the increase in nutrients led to an increase in zooplankton, but not in the phytoplankton. The phytoplankton, in fact, decreased slightly. The ungrazed periphyton also increased (a in the figure). With *Rana* present, the results are different (b in the figure). With low nutrients, *Rana* grazed the periphyton to low levels. Increased nutrients led to increased biomass of both periphyton and *Rana*. The periphyton did not reach the high biomass seen without *Rana*, while the competing phytoplankton reached its highest biomass. The zooplankton at high nutrient levels were less dense than when *Rana* was absent. Indeed, with *Daphnia* absent, the zooplankton were only slightly more abundant in the high nutrient treatment. They were, however, substantially more abundant in the treatments in which *Daphnia* and *Rana* were present.

Without *Rana*, the results fit the Lotka-Volterra predictions for increasing nutrients. The periphyton was not grazed and it increased; the phytoplankton was grazed and it did not increase, but the zooplankton did. With *Rana* present, both plant groups increased, one herbivore, *Rana*, increased, while

1. Leibold, M. A. & Wilbur, H. M. *Nature* **360**, 341–343 (1992).
2. Martin, J. H., Gordon, R. M. & Fitzwater, S. E. *Nature* **345**, 156–158 (1990).
3. Pimm, S. L. *Food Webs* (Chapman & Hall, London, 1982).
4. McNaughton, S. J., Oesterheld, M., Frank, D. A. & Williams, K. J. *Nature* **341**, 142–144 (1989).
5. Moen, J. & Oksanen, L. *Nature* **353**, 510 (1992).
6. Cyr, H. & Pace, M. L. *Nature* (in the press).
7. Oksanen, L., Fretwell, S. D., Arruda, J. & Niemelä, P. *Am. Nat.* **118**, 240–261 (1981).
8. Arditi, R. & Ginzburg, L. R. *J. theor. Biol.* **139**, 311–326 (1989).
9. Pimm, S. L. *The Balance of Nature? Ecological Issues in the Conservation of Species and Communities* (University of Chicago Press, 1991).

the herbivorous zooplankton either did or did not, depending on its composition.

The conclusion is simple. The one-species-per-trophic-level models work when the systems are indeed that simple. Real ecosystems of course contain many interconnected food chains — but when the number of chains is as small as two,

the effects of nutrient enrichment depend at least in part on the structure of the food web<sup>9</sup>. □

Stuart L. Pimm is in the Department of Zoology and Graduate Program in Ecology, University of Tennessee, Knoxville, Tennessee 39776, USA.

## NUCLEAR PHYSICS

# Invisible radioactivity

W. R. Phillips

NUCLEAR  $\beta$ -decay is best known as an event in which an electron is created and ejected from a nucleus at high speed. But physicists at the German heavy-ion laboratory GSI, near Darmstadt, have succeeded in observing an exceedingly rare variant in which the electron produced is born into a bound atomic orbital<sup>1</sup>. One potential by-product of this discovery is the prospect of making an accurate measurement of the mass of the electron neutrino.

The electron created in  $\beta$ -decay, in which a neutron is transmuted into a proton, typically emerges from the nucleus as a freely moving particle (with a continuum of possible kinetic energies), but there is always, in principle, the possibility that it will be created in a bound, unoccupied atomic orbital of the daughter ion. Bound-state  $\beta$ -decay should compete best when the maximum energy available to the electron is small; but this condition also causes the lifetime of the emitter to be long. For example, <sup>187</sup>Re, which  $\beta$ -decays to <sup>187</sup>Os with a maximum energy just under 3 kiloelectronvolts, has a radioactive half-life of  $4 \times 10^{10}$  years, a truly cosmological timescale. Even in this decay, which can be used as a cosmochronometer<sup>2</sup> to determine the age of our Galaxy, the fraction of bound-state decays may be no more than 1 per cent (ref. 3).

The chances of bound-state decay are considerably greater if the decaying atom is highly ionized: this leaves tightly bound inner electron shells unoccupied and available to the emerging electron, and these shells have a relatively large binding energy to be gained by the decay process. This opens up the intriguing prospect that nuclei exist that are stable in neutral atoms but which can undergo bound-state  $\beta$ -decay when in highly charged or bare ions (continuum decay would still be impossible). One such nucleus is <sup>163</sup>Dy, which in the neutral atom is stable against  $\beta$ -decay by a few kiloelectronvolts, but can undergo bound-state decay into the K or L atomic shells of <sup>163</sup>Ho when highly ionized. This presented the Darmstadt workers with an easier way to search for

bound-state  $\beta$ -decay. One can start with a number of fully stripped <sup>163</sup>Dy nuclei and count the number of single-electron <sup>163</sup>Ho ions produced in a given time. All the <sup>163</sup>Ho ions must have resulted from bound-state decay, and their number is a measure of the decay probability per unit time.

The experiment was made possible by the recent construction of the heavy-ion accelerator and storage ring at Darmstadt. Up to 100 million partially ionized <sup>163</sup>Dy ions were accelerated in a heavy-ion synchrotron to an energy of 47.9 gigaelectronvolts, when they are moving at about two-thirds the speed of light. At this speed most Dy ions, after passing through a thin foil, have all their electrons stripped off. In this way bare <sup>163</sup>Dy ions were produced and injected into the GSI storage ring. The highly evacuated ring's magnetic elements constrained the bare nuclei to move stably around the ring orbit. Ions of the same speed and the same mass-to-charge ratio move on similar orbits, and the ring can store ions as heavy as <sup>163</sup>Dy for periods of up to a few hours without substantial loss, during which time a few could decay to give <sup>163</sup>Ho ions.

Because these daughter ions, with their one electron, have the same mass-to-charge as the parent ions, they also remain in stable orbit around the ring. By passing the ion beam through a gas jet after a short accumulation time, the Darmstadt team could strip the final electron from the Ho ions, which could then be extracted from the storage ring and counted. Suitably adjusted for the loss of Dy ions from the beam, the slow count rate for the appearance of Ho ions translates into a half-life for the bound-state  $\beta$ -decay of  $47^{+5}_{-4}$  days, close to the predicted value<sup>4</sup> of 50 days.

One interesting outcome of the experiment is the first good upper limit on the mass of the electron neutrino (better-known experiments on the 'neutrino' mass in fact deal with the antineutrino). Bound-state  $\beta$ -decay (accompanied by the ejection of an antineutrino) is the inverse process of orbital electron capture, in which a nucleus reduces its

charge by capturing an electron from an atomic orbital (simultaneously creating and ejecting a neutrino). A <sup>163</sup>Ho nucleus in a neutral atom decays to <sup>163</sup>Dy by capturing an electron from the atomic M shell or from outer shells with less binding energy. A bare <sup>163</sup>Dy nucleus decays into the K shell around a <sup>163</sup>Ho nucleus. The ratio of probabilities for the <sup>163</sup>Dy decay and for electron capture by <sup>163</sup>Ho depends on electron densities at the nuclei and on other factors known reasonably accurately. But it also depends on the phase (or momentum) space available to the product antineutrino or neutrino which, in turn, depends on the particle's energy. This is determined by the exact mass difference between <sup>163</sup>Dy and <sup>163</sup>Ho nuclei and also by the masses of the antineutrino and neutrino, if they have mass.

Comparison of the known electron-capture rates for <sup>163</sup>Ho decay with the decay rate measured in the GSI experiments gives independent expressions involving these separate masses. Neglecting the mass of electron antineutrino, for which the present<sup>5</sup> upper limit is 17 electronvolts, enables limits to be placed on the mass difference between the nuclei and on the mass of the electron neutrino. The latter is less than 410 electronvolts at the 68 per cent confidence level. (There are other measures of the neutrino mass which give lower limits, but relying on another exotic radioactive process — neutrinoless double  $\beta$ -decay — which has yet to be observed, their result depends on the character of the neutrino. The non-observation of this process gives a mass<sup>6,7</sup> only if the neutrino is its own antiparticle, as suggested once by Majorana. The more usual assumption, dating back to Dirac, is that the neutrino and antineutrino are distinct particles.)

Although the upper mass limit looks poor by comparison with the antineutrino mass determinations, improvements merely require better measures of the respective radioactive decay rates and of the precise mass difference between <sup>163</sup>Dy and <sup>163</sup>Ho nuclei. In detecting bound-state  $\beta$ -decay, the GSI workers have already vaulted the most formidable hurdle. □

W. R. Phillips is in the Department of Physics, University of Manchester, Oxford Road, Manchester M13 9PL, UK.

1. Jung, M. et al. *Phys. Rev. Lett.* **69**, 2164–2167 (1992).
2. Clayton, D. D. *Astrophys. J.* **139**, 637–663 (1964).
3. Chen, Z., Rosenberg, L. & Spruch, L. *Phys. Rev. A* **35**, 1981–1990 (1987).
4. Takahashi, K., Boyd, R. N., Mathews, G. J. & Yokoi, K. *Phys. Rev. C* **36**, 1522–1528 (1987).
5. Bertsch, G. F. (ed.) *Review of Particle Properties*, *Phys. Lett.* **B239**, 1–516 (1990).
6. Witherell, M. S. *Nature* **331**, 392–393 (1988).
7. Hykawy, J. G. et al. *Phys. Rev. Lett.* **67**, 1708–1711 (1991).



# Deconstructing the MHC

Peter Parham

CRYSTALLOGRAPHIC structures of class I major histocompatibility molecules now come thick and fast. Three recently unveiled<sup>1-3</sup> structures of a mouse class I molecule reconstituted with synthetic peptides show how peptides of different length and sequence bind with similar conformation: an extended  $\beta$ -structure, pinned down at the ends but with freedom in the middle. Papers on pages 364 and 367 of this issue<sup>4,5</sup> and in *Cell*<sup>6</sup> continue this theme for the human class I molecules HLA-A68 and HLA-B27. The principle to emerge, namely that peptides selected for class I binding have conformationally constant and variable

perusers of *Nature* that the class I MHC molecule is an intrinsically female structure (Fig. 1). The shapes of the bound peptides remained decidedly uncertain, however, and it is to resolution of this issue that a cavalry of crystallographic gentlemen have charged.

With refinement of HLA-A2 and solution of a second molecule, HLA-A68, substructure (pockets A through to F) was assigned to the peptide-binding groove, and regions of the extra electron density that were recognizably peptide<sup>9,10</sup> were identified. In a third molecule, HLA-B27, the peptides were better behaved and gave a connected pathway of clear density due to their common conformation — an extended  $\beta$ -structure with a kink between residues 3 and 4 (refs 6,11). The HLA-A68 structure has now been refined and Guo *et al.*<sup>4</sup> confirm that the bound peptides are heterogeneous compared with those in the B27 crystals. These authors retrieved the first three and the last two peptide residues, the conformation of which is like that of the homologous residues in the B27 peptides, but cannot connect the two groups. Correlated with this difference are sequences of endogenous peptides showing that A68-bound peptides are variable in length, 8–10 residues against the nonamers extracted from B27.

For some time, H-2K<sup>b</sup>, a mouse class I molecule, has been known to bind peptides of different length, and from analysis of its structure<sup>1-3</sup> we see how they fit into the same peptide-binding groove (Fig. 2). In these studies, crystallography was performed not on molecules containing a mixture of natural peptides (as done for HLA-A2, A68 and B27) but on products of cloned genes deliberately loaded with synthetic peptides. The loaded peptides were an octamer — RGYVYQGL — from vesicular stomatitis virus and a nonamer — FAPGNYPAL — from Sendai virus. Again, both peptides assume an extended  $\beta$ -conformation in which the first three and the final four residues of the peptide interact similarly with pockets of the H-2K<sup>b</sup> groove. The difference in length between the two peptides occurs at the midriff, with glycine 4 and asparagine 5 of the Sendai peptide protruding from the binding site, as a  $\beta$ -bulge

roughly in the same position as valine 4 of the vesicular stomatitis virus peptide<sup>1</sup>.

Using an analogous approach, Silver *et al.*<sup>5</sup> reconstituted HLA-A68 with a nonameric influenza peptide and now report its structure. Despite its very different sequence (KTGGPIYKR) the conformation of this bound peptide is essentially the same as the mixture of B27-binding peptides and very similar to that of the Sendai nonamer in H-2K<sup>b</sup>. That a defined peptide binds to A68 with a preferred structure suggests, following the H-2K<sup>b</sup> precedent, that the absence of connectivity between the first three and the last two residues in the mixed A68-binding peptides of Guo *et al.*<sup>4</sup> is due to differences in peptide length affecting the severity of the bulge, and thus of the local conformation within the midriff.

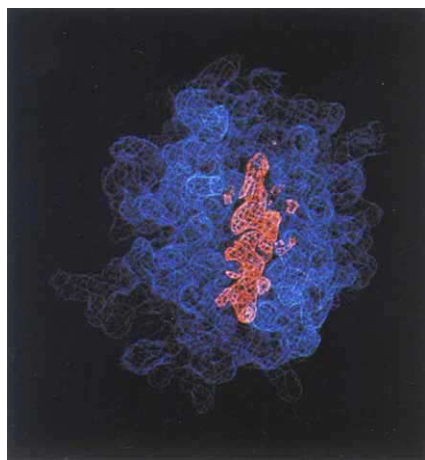


FIG. 1 Van de Waal's surface representation of the top of the HLA-A2 molecule (blue) showing the bound peptides (pink). From ref. 7.

regions, explains how class I molecules are promiscuous with regard to the peptides they bind yet paradoxically seem to be well adapted to any particular partner.

Throughout the 1970s crystallography of class I molecules was much discussed over port and cigars, but remained a fantasy until Pamela Bjorkman obtained crystals of HLA-A2 in 1979. Eight years passed before a sampling of their secrets<sup>7,8</sup>. But then it suddenly became obvious — to both structural biologists and immunologists — that three-dimensional structure had much to contribute to a growing subspeciality, the study of antigen presentation. HLA-A2, a noble structure, transcended its loops, sheets and helices to tell us, after due reference to the immunological encyclopaedia, all about its function. A deep groove formed by polymorphic parts of the class I heavy chain was seen to be filled with blurry 'extra electron density' coming from a mix of intracellularly bound peptides. From these first images it was apparent to seminar goers and

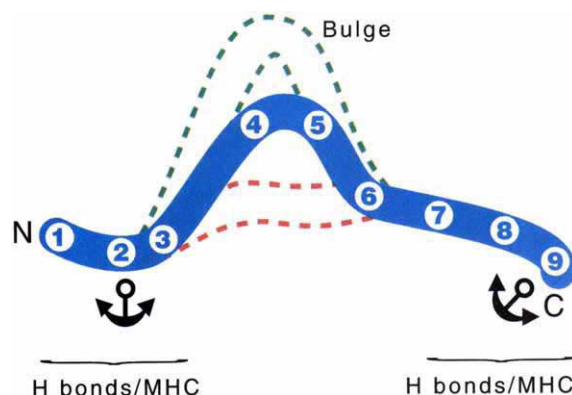


FIG. 2 Diagram of the conformation of a nonamer peptide in the class I MHC binding groove. The less bulging path taken by an octamer in H-2K<sup>b</sup> is shown in red. Hypothetical bulging by longer peptides is shown in green. Anchor residues for the A68 binding peptides are indicated.

These papers echo each other in pronouncing general principles for peptide interactions with class I MHC molecules. Peptides bind in the same orientation. Their amino and carboxy termini are held within conserved pockets at the ends of the binding groove by networks of hydrogen bonds which provide a common foundation for the strong, practically irreversible, binding that characterizes the interaction between class I molecules and peptide. Augmenting the conserved

1. Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. *Science* **257**, 919–927 (1992).
2. Matsumura, M., Fremont, D. H., Peterson, P. A. & Wilson, I. A. *Science* **257**, 927–934 (1992).
3. Zhang, W., Young, A. C. M., Imarai, M., Nathenson, S. G. & Sacchettini, J. C. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8403–8407 (1992).
4. Guo, H.-C. *et al. Nature* **360**, 364–366 (1992).
5. Silver, M. L., Guo, H.-C., Strominger, J. L. & Wiley, D. C. *Nature* **360**, 367–369 (1992).
6. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Cell* **70**, 1035–1048 (1992).
7. Bjorkman, P. J. *et al. Nature* **329**, 506–512 (1987).
8. Bjorkman, P. J. *et al. Nature* **329**, 512–518 (1987).
9. Saper, M. A., Bjorkman, P. J. & Wiley, D. C. *J. molec. Biol.* **219**, 277–319 (1991).
10. Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L. & Wiley, D. C. *Nature* **342**, 692–696 (1989).
11. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Nature* **353**, 321–325 (1991).

interactions are variable ones, between one or two anchoring peptide side chains and complementary polymorphic pockets of the class I groove. The middle of the peptide is less constrained by MHC and more exposed for direct interaction with T-cell receptors, and differences in length and sequence are accommodated by bulging away from the groove. Most binding interactions with the peptide involve main-chain atoms and not those of the side chains. Thus the contribution of peptide sequence to high-affinity binding is centred on the anchor residues, leaving other positions open to variation. In this manner, many peptides of different sequence can bind with similar affinity to a given class I MHC molecule, so providing the passport for promiscuity.

Protein crystallography has historically been served by a gentlemen's agreement whereby short papers announcing new

crystals serve to dissuade the competition. Tradition is going by the board as the field embeds itself into every area of biomedical research, as illustrated by the independent solution of H-2K<sup>b</sup> bound to the Sendai virus octamer<sup>1,3</sup>. Such congruence, however, reveals the relative ease of solving new class I structures once the tree of knowledge has already been tapped. This autumn's crystallographic bonanza undoubtedly establishes prototypic structures for peptides bound to class I MHC molecules. They seem to be highly constrained at the extremities but have a tendency to bulge in the middle — characteristics of the male, don't you think? □

Peter Parham is in the Departments of Cell Biology, and Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305, USA.

## NEUROBIOLOGY

# Elements of visual perception

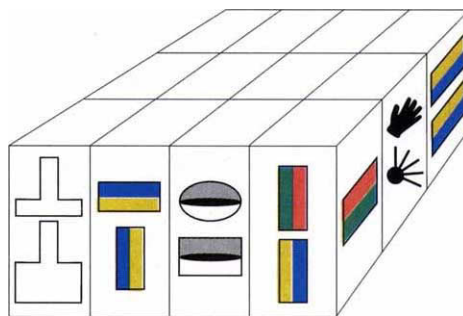
Michael P. Stryker

How do we recognize and remember complex visual objects? Experiments on monkeys by Fujita *et al.*, described on page 343 of this issue<sup>1</sup>, suggest that a moderate-sized collection of almost iconic figures may be the alphabet in which our visual memories are written.

These authors have long studied the response properties of neurons in the anterior inferotemporal cortex (IT), the part of the brain necessary for visual object recognition and most closely connected to neural circuits thought to subserve memory<sup>2</sup>. It is in the IT that neurons were discovered long ago to respond selectively to images of complex objects, such as a face or a hand<sup>3</sup>. In several laboratories over the years, images from many trays of slides — of people, even of scenes from holidays on the farm or at the beach — have been shown to monkeys, and each IT cell responded well to different subsets of these stimuli. These findings engendered several interpretations, including the notion of the 'grandmother cell' — that you are able to recognize your grandmother because somewhere in the IT there is a cell selective for her image. The greatest difficulty with this notion is testing it. Because one could try only a limited array of stimuli, one could never be confident one had found the best stimulus for a cell and never be certain how selective such cells really are.

Tanaka's laboratory has brought a re-

ductionist approach<sup>4</sup> to the study of the IT. Starting with a large collection of toy animals, vegetables and natural objects as visual stimuli, they have progressively



Columnar organization of the IT cortex. Columns of cells selective for a range of similar stimuli (for example, the coloured bars) are interspersed among columns unresponsive to those stimuli but selective for other complex visual qualities.

simplified and abstracted the features necessary for eliciting the strongest responses of single neurons, often improving the response over the best obtained with any complex real object. Their new finding is that these visual patterns organize the IT into repeated arrays of columns. In areas of the brain whose properties we understand relatively well, such as the primary visual cortex, cortical columns organize the most significant features of neural responses. The fact that the iconic stimuli described by Fujita *et al.* also determine the columnar arrangement of the IT is powerful testimony to the validity of their characterization of the cells' receptive fields.

Why should the most important stimuli for an area of cortex organize its columns? One possibility is that perception heeds the combined output signal from a series of neighbouring columns. This approach has led to explanations of a variety of adaptation phenomena (that is, changes in perception following prolonged exposure to one stimulus<sup>5</sup>). It also allows distance in the cortex to be a surrogate for distance in the representational space, and for neural computations that depend on the computational geometry of cortical maps<sup>6</sup>.

Another possible reason is developmental — that neuronal connections are shaped by correlations in activity, and that the most important stimuli, which cause the greatest modulations of activity, come to be the most closely linked<sup>7,8</sup>. Often, but not inevitably, these close synaptic links would take the form of columns, although in other circumstances the same process aggregates cells with similar responses into patches, clumps or laminae instead. Regardless of the reason, the fact is that columns commonly specify the most significant stimulus features.

A columnar encoding scheme is quite the opposite of that proposed elsewhere for this area of the brain<sup>9,10</sup> as well as for the olfactory system<sup>11</sup>. There (outside the pheromone-detecting system at least), each cell's response has been conceived of as one dimension of a very high-dimensional space in which stimuli are classified by some sort of neural network. It is thus advantageous for the cell responses to be as different from one another as possible, so that the dimensions of the classification space are maximally orthogonal; that is, the more distinct each cell's response is from the response of any other cell, the more information is carried by that cell. The view that every cell contributes its unique response spectrum to perception makes a virtue of necessity when the experimenter does not discern the simplicity of the properties that organize a brain area. Like the notion of the grandmother cell, it is difficult to test. It will

1. Fujita, I., Tanaka, K., Ito, M. & Cheng, K. *Nature* **360**, 343–346 (1992).
2. Iwai, E. & Mishkin, M. *Vision, Memory, and the Temporal Lobe* (Elsevier, New York, 1990).
3. Bruce, C., Desimone, R. & Gross, C. J. *Neurophysiol.* **46**, 369–384 (1981).
4. Tanaka, K. *et al.* *J. Neurophysiol.* **66**, 170–189 (1991).
5. Blakemore, C. *et al.* *Nature* **228**, 37–39 (1970).
6. Schwartz, E. L. *Computational Neuroscience* (MIT Press, Cambridge, Massachusetts, 1990).
7. Zaksas, K. R. & Stryker, M. P. *J. Neurophysiol.* **59**, 1410–1429 (1988).
8. Purves, D., Riddle, D. R. & LaMantia, A. S. *Trends Neurosci.* **15**, 362–368 (1992).
9. Rolls, E. T. in *The Neural and Molecular Bases of Learning* (eds Changeux, J. P. & Konishi, M.) 503–540 (Wiley, Chichester, 1987).
10. Young, M. P. & Yamani, S. *Science* **69**, 5008 (1992).
11. Haberly, L. B. & Bower, J. M. *Trends Neurosci.* **12**, 258–264 (1989).
12. Karni, A. & Sagi, D. *Proc. natn. Acad. Sci. U.S.A.* **88**, 4966–4970 (1991).



be interesting to see whether our present views of orthogonalized representations in, for example, the olfactory system, stand the test of time.

If the IT columns function as suggested, the size of the anterior IT and the size and spacing of the IT columns of similar specificity set an upper limit of perhaps a thousand for the number of different 'letters' in the visual alphabet. Given the remarkable capacity for perceptual learning<sup>12</sup>, one may ask how malleable these columns and receptive

fields are. Are they dramatically different in animals reared in different early visual environments? Does a period of intense experience with particular stimuli make them change even in adults? These are among the next questions that will have to be addressed. □

Michael P. Stryker is in the W. M. Keck Foundation Center for Integrative Neuroscience, Department of Physiology, University of California, San Francisco, California 94143-0444, USA.

## ASTRONOMY

# Fast process, slow progress

M. G. Edmunds

TWENTY-FIVE years should be a long time in science, yet that many years have passed since Lynden-Bell proposed 'violent relaxation' as a mechanism to explain the internal structure of galaxies<sup>1</sup> and still no really satisfactory description of the underlying physics of the process has been put forward. A new prescription by Spergel and Hernquist<sup>2</sup>, redefining the way one accounts for the entropy of a galaxy's constituent stars, may at last show the way forward.

The problem is one of timing. Violent relaxation (an apparent contradiction in terms) is meant to explain how a galaxy can share energy among its stars to reach some sort of equilibrium state. A gas equilibrates through energy-exchanging collisions between its particles, but the equivalent galactic process — collisions between stars — is far too slow. The average stellar density is too low for encounters to occur often enough, and the galaxy's stars remain effectively collisionless over times far longer than the age of the Universe.

The essence of violent relaxation is a rapid change in the overall shape or size of the system — either by its initial gravitational collapse from a less dense state, or through gravitational interaction with another galaxy. This leads to a sharing around of energy with the stars scattering off the changing global gravitational field or off clumps of stars within it. The timescales of such processes are much shorter.

Violent relaxation certainly happens. Countless numerical models allowing gravitational interaction of a large number of stars have shown a rapid approach to a smooth structure whose projected surface density fall-off with radius, known as the  $r^{1/4}$  law, resembles that of real elliptical galaxies. The persistent problem is why does this particular structure arise? The structure does not represent a state of maximum entropy. Indeed, the maximum-entropy structure

for a self-gravitating system is a rather strange one, with a tightly bound central core of stars surrounded by a diffuse (and infinite) halo. On a very long time-scale, the two-body interactions between stars in a galaxy will presumably lead to this structure, but this takes far too long to be of interest. What is it that rapidly moulds the galaxy into a regular structure and holds it there until its ultimate relaxation to a maximum-entropy state? There must be some sort of quasi-maximized entropy state on the way.

Last year, Hjorth and Madsen suggested<sup>3</sup> that the appearance of  $r^{1/4}$ -law behaviour did not depend on the details of the initial conditions before the galaxy formed, but would always arise if the galactic gravitational potential well were sufficiently deep in the centre — in other words, if the central density were sufficiently high. This is a bit like a man falling down a lift shaft: once he has fallen a few feet the result is going to be pretty much the same, however deep the shaft. But this still does not really provide a very clear reason for the particular observed structure, although it does suggest that one is always going to get roughly the same outcome.

Emphasizing that it is not the maximum-entropy state that is reached, Funato *et al.* argue<sup>4</sup> that violent relaxation does not even try to reach thermal equilibrium. They suggest, from numerical models of head-on collisions of pairs of spherical galaxies, that the rapid change in gravitational potential (which is what always drives the violent relaxation) effectively increases the energy of nearly all the stars. But this sharing out of energy seems to be very inequitable: the stars with the largest energy to begin with gain more than those with low energy. Although this sort of behaviour is all too familiar in a financial context, it is the opposite of what is required to establish thermal equilibrium. But it does imply that the violently relaxed

galaxy could retain much of the structure of its pre-relaxed state, such as a particular radial distribution of stellar age or metallicity. This property is not unique to inequitable energy gain, however, as it could arise in any energy-sharing process that does not continue to completion (violent relaxation certainly halts when the gravitational potential stops changing).

Some real progress in elucidating the basic process seems to have been made by Spergel and Hernquist, who argue<sup>2</sup> that a combination of two physical constraints can lead to the  $r^{1/4}$  structure. First, the interaction that an individual star experiences during the rapid violent-relaxation phase feels like a series of discrete scatterings — presumably off transient clumps of other stars. Second, the final state of the scattered star depend on its initial state: quantities such as its angular momentum and energy are conserved apart from any addition or decrease due to the impulse of the scattering. This second constraint restricts the phase space into which the system can evolve.

The unsatisfactory maximum-entropy state is reached only if the system can access the whole of phase space. With the new constraints, it is possible to define a quantity that looks like entropy, but which includes a density of 'accessible' states instead of a global density of states. This new form — a sort of achievable entropy — is what seems to be maximized in the violent relaxation and results in the characteristic  $r^{1/4}$  structure. The definition of the correct density of accessible states is not trivial, and the authors apparently had to ignore the observed preservation of anisotropy in a violently relaxing collapse. But this does seem to be a promising avenue to pursue.

There may, of course, be more to real galaxies than just stellar dynamics. The stars must have formed out of gas at some time, and dissipational processes in this gas could still be critical in setting up the structure of elliptical galaxies. A better understanding of the physics of violent relaxation will surely be of benefit in sorting out the relative importance of dissipational (gas dynamics) and dissipationless (stellar dynamics) processes in galaxy formation. □

M. G. Edmunds is in the Department of Physics and Astronomy, University of Wales, College of Cardiff, PO Box 913, Cardiff CF2 3YB, UK.

1. Lynden-Bell, D. *Mon. Not. R. Astr. Soc.* **136**, 101–121 (1967).
2. Spergel, D. N. & Hernquist, L. *Astrophys. J.* **397**, L75–L78 (1992).
3. Hjorth, J. & Madsen, J. *Mon. Not. R. Astr. Soc.* **253**, 703–709 (1991).
4. Funato, Y., Makino, J. & Ebisuzaki, T. *Pub. Astr. Soc. Japan*, **44**, 291–301 (1992).

# Still waters run still deeper

Christopher M. Sorensen

WATER expands upon freezing (one of only a handful of substances to do so), its liquid state has a temperature of maximum density (along with only  $\text{SiO}_2$ ), uniquely when cold its viscosity decreases upon addition of pressure, and it has anomalously high melting and boiling points, specific heat and dielectric constant. Water and its solutions are the most far reaching of solvents. And, of course, it is the grand mediator of the biochemistry that allows us to live and so to contemplate its marvels. The challenge is to explain its macroscopic nature from the microscopic character of the single molecule. Rising to this challenge on page 324 of this issue<sup>1</sup>, Poole *et al.* not only propose a new explanation for water's wonders, but deepen them too.

Qualitatively, the source of water's properties can be traced to two fundamental properties of the molecule: its ability to self associate through hydrogen bonding; and its space-filling tetrahedral geometry. Comparison with the hydrides of oxygen's neighbours in the periodic table readily demonstrates this.  $\text{H}_2\text{S}$ ,  $\text{H}_2\text{Se}$  and  $\text{H}_2\text{Te}$  are tetrahedral but do not hydrogen bond well; they are relatively simple liquids with low freezing and boiling points.  $\text{NH}_3$  and  $\text{HF}$  form good hydrogen bonds, but the associated structure in the liquid is not space filling.  $\text{H}_2\text{O}$ , however, can both accept and donate two protons, giving a total of four hydrogen bonds to form a tetrahedral arrangement which allows the condensed phases to fill space with a low-density, four-coordinated mesh of molecules. With this microscopic picture, many of the properties of water can be qualitatively explained. For instance, as the temperature falls, weak hydrogen bonds stabilize, so forcing the molecules apart into the fourfold arrangement. This is why ice is less dense than liquid water, and shows that the liquid temperature of maximum density is the anticipation of the solid phase.

So it was until the early 1970s, when some people began asking 'what if' one were to supercool water below its freezing point. In a classic paper, Speedy and Angell<sup>2</sup> showed that a wide variety of thermodynamic and transport properties of water all appear to diverge as the temperature is lowered into the supercooled regime (Fig. 1). Remarkably, the extrapolated point of divergence was at  $-45^\circ\text{C}$  for every parameter studied.

Explanations were proposed. All invoked large fluctuations in density and entropy caused by the self-associating nature of water. These fluctuations are the fundamental cause of water's special

heat capacity, expansivity and compressibility. But why should they come about? One reasonably successful proposal, the stability-limit conjecture<sup>3</sup>, was that supercooled water approaches a spinodal, the metastable limit line of second-order phase transitions beyond which no physically realizable states exist, and along which fluctuations and their concomitant susceptibilities diverge (Fig. 2). Furthermore, the conjecture proposed that the lower branch of the liquid-vapour spinodal (curve CS in Fig. 2) turns around to appear in the metastable region of the liquid-solid transition. Supercooling the liquid would bring one close to this spinodal, which would be at  $-45^\circ\text{C}$  for one atmosphere of pressure, so giving the divergent behaviour observed.

Poole *et al.* have simulated water on a computer. The advantage is that meta-

stable regimes can be explored in depth without the real-world limitation of nucleation from the liquid to the solid state. The disadvantage is, of course, not knowing how well the computer repre-

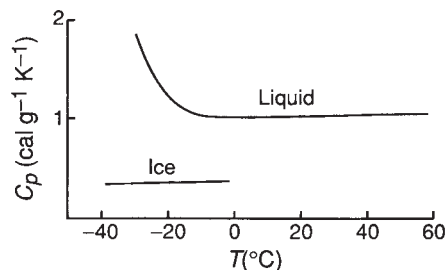


FIG. 1 Divergence in water. Specific-heat measurements at constant pressure as a function of temperature<sup>5</sup> reveal strong changes below  $-20^\circ\text{C}$ .

sents real water. Poole *et al.* find that for computer water, at least, the 're-entrant' behaviour of the spinodal does not occur — the stability-limit conjecture does not hold. So what is the source of the anomalous supercooled behaviour? A

## Nuclear age reaches half century

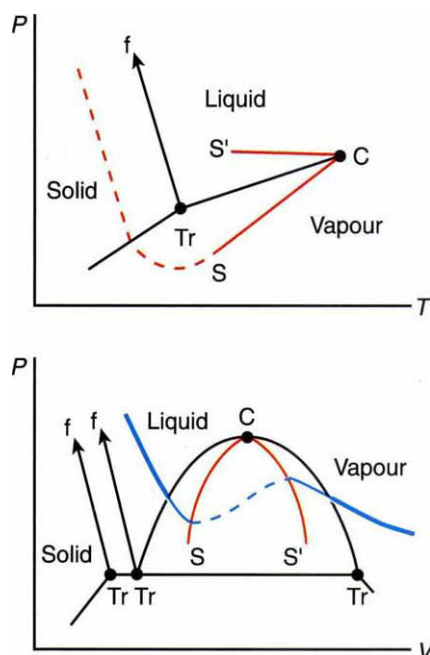
IMAGE  
UNAVAILABLE  
FOR COPYRIGHT  
REASONS

ON 2 December, the nuclear world will have been with us for 50 years. Shortly before 3.30 p.m. on that day in 1942, Enrico Fermi instructed George Weil to withdraw a manual cadmium control rod a crucial 12 inches from their experimental nuclear pile, and watched the neutron flux steadily increase. Within minutes it became clear that the reactor had reached criticality — the nuclear fission reactions were self-sustaining. The moment was later recaptured in this painting by Gary Sheahan, showing Fermi (bald, to the rear, seated before the instrument panel), Weil (below, by the pile), the "suicide squad" (right corner) — delegated to pour cadmium sulphate solution over the pile if all went wrong — and 37 other members of the Manhattan Project. The construction of the pile, comprising 400 tons of graphite bricks and 50 tons of uranium metal and uranium oxide, had started just over two weeks earlier in the now renowned squash court beneath the west stands of the University of Chicago's football stadium. The pile completed, the experiments started on the morning of 2 December, with Weil withdrawing the control rod 6 inches at a time while the reactor characteristics were checked. The scientists returned after lunch, certain of reaching criticality, which they duly did. The reactor was allowed to run for half an hour before Fermi called a halt. Eugene Wigner (seated, facing us) produced a bottle of chianti which he had been concealing all day, and the experiment was toasted. Fermi later commented on the lack of drama that day: "no fuses burned, no lights flashed. But to us it meant that release of atomic energy on a large scale would be only a matter of time".

R.P.



FIG. 2 The concept of a spinodal comes from the van der Waals theory of the liquid-vapour transition. This figure shows two perspectives of a van der Waals gas-phase diagram with a liquid-solid transition ( $f$ ) added. Black rules mark phase boundaries. In the pressure-volume ( $P$ - $V$ ) perspective, the inverted parabola marks the liquid-vapour coexistence line; depressurizing the liquid brings the system down an isotherm (solid blue line) to the coexistence line, at which point the liquid will boil, and the curve continues on the right of the diagram. (C marks the critical point, where liquid turns to vapour continuously.) The fine blue line marks the metastable state. Points of zero slope for the isotherms correspond to infinite compressibility. These points are on the spinodal (red), the limit of stability (so that the broken blue line is inaccessible). In the pressure-temperature ( $P$ - $T$ ) perspective, these spinodal lines project above and below the liquid-vapour coexistence line. The stability-limit conjecture is that the spinodal, when extended to lower temperature, curves up (broken curve) around the triple point Tr, missing the solid-vapour coexistence curve, to run roughly parallel to the freezing line,  $f$ .



new second-order transition in the super-cooled regime. A critical point between the two known phases of amorphous ice, the high- and low-density phases, would do the trick. Although, like the spinodal, this critical point could not be reached experimentally, its associated fluctuations would influence metastable states in its vicinity. (I have yet to hear of another material in which two solid phases could have a critical point — another plus for water. The low symmetry of the two amorphous phases allows this possibility.)

The stability-limit conjecture should not be passed off lightly. The conjecture does a remarkable job of correlating an extensive set of thermodynamic data. An attractive feature of the conjecture is in the interplay of the liquid temperature of maximum density (TMD) and the spinodal. As pressure is decreased, the TMD increases, inscribing a line on the phase diagram. One can prove a thermodynamic necessity that the TMD line, if it intersects the spinodal, will do so at a point where the spinodal has zero slope. It is known that the TMD line has negative slope in the pressure-temperature phase diagram and reasonable extrapolation implies an intersection with the spinodal. Hence the spinodal having a positive slope (curve CS, Fig. 2) at high temperature would pass through zero slope at the intersection and then presumably have a negative slope at lower temperature, causing it to be re-entrant.

Experimentally, Angell and co-workers<sup>4</sup> trapped water in microscopic inclusions in quartz. With proper thermal processing, they achieved extreme tension — that is, negative pressure — in the trapped water and mapped

out the homogeneous nucleation curve for water near the spinodal curve. Although the re-entrant spinodal could not be observed directly, they did establish the validity of an equation of state extrapolated far from its data base, an equation that predicts a re-entrant spinodal so supporting the stability-limit conjecture. For now, perhaps, it's a matter of taste: do you prefer extrapolations or simulations?

Given the results of Poole *et al.*, what should be our physical picture of water? Certainly, self-association and low-density tetrahedral arrangements are still the key microscopic properties. But deep in the experimentally inaccessible super-cooled regime, the two known phases of amorphous ice can coexist, just as liquid and vapour can coexist in equilibrium regimes. And this coexistence is terminated at a critical point where the two amorphous phases become identical, and where large fluctuations and the concomitant critical phenomena of divergent thermodynamic and transport properties occur. Although never attainable in a real experiment, this distant critical point would send out its influence into the attainable regimes, perturbing the thermodynamic landscape and creating a most remarkable liquid. □

Christopher M. Sorensen is at the Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA.

1. Poole, P. H., Sciortino, F., Essman, U. & Stanley, H. E. *Nature* **360**, 324–328 (1992).
2. Speedy, R. J. & Angell, C. A. *J. chem. Phys.* **65**, 851–858 (1976).
3. Speedy, R. J. *J. phys. Chem.* **86**, 982–991; 3002–3005 (1982).
4. Zheng, Q., Durben, D. J., Wolfe, G. H. & Angell, C. A. *Science* **254**, 829–832 (1991).
5. Angell, C. A., Shuppert, J. & Tucker, J. C. *J. phys. Chem.* **77**, 3092–3099 (1973).

## Hyperfine print

THE human eye is ill-adapted to sustained reading. Some of us have very bad close vision anyway, and we all get long-sighted as age advances. Reading glasses are sold in huge numbers. Daedalus has been looking at the problem from the other end. What sort of print, he wonders, would look sharp and clear when seen out of focus? In principle, the answer is simple. A black letter on a white background would need its outside edge to be blacker than black, while the paper just bordering the letter would have to be whiter than white. Defocused imaging would then reduce this heightened contrast back to normal. The letter would appear a consistent black on a consistent white background.

A whiter-than-white border presents no great problem. Many fluorescent pigments absorb invisible ultraviolet and reradiate its energy in the visible spectrum, thus seeming brighter than the incident light. A well-judged mixture of such fluorescent pigments should give a neutral hyper-white. Blacker-than-black seems a more serious challenge. But Daedalus points out that all conventional black pigments reflect a few per cent of light from their surface. A black pigment that reflected, say, 0.01 per cent of the light, would bear the same relationship to ordinary black ink as that black does to white paper.

This hyper-black ink will exploit the principle of the anechoic chamber. Such a chamber is lined with many protruding fingers of sound-damping material. Impinging sound is repeatedly reflected between them until it is totally absorbed. A surface anechoic to light would need much finer fibres, like a sort of small-scale black velvet (whose blackness derives from just this effect). DREADCO's pigment technologists are seeking an ink that dries in the form of innumerable protruding black micro-fibres — the ultimate in crackle-finish paints. It may not be too difficult; such surfaces are formed spontaneously by magnetic liquids in a field applied normal to the surface.

When perfected, DREADCO's 'Autofocus Print' will be wonderfully easy on the imperfect eye. Thanks to the hyper-black borders of the letters, and the hyper-white paper around them, it will seem sharp and clear to the most long-sighted reader. Normal readers will see it just as clearly. Focusing by automatic reflex on the most intelligible image, they will simply focus in front of or behind the paper. The amazing result will be standard black-on-white letters, apparently suspended in space. Daedalus intended his invention for the elderly and distinguished readership of scientific journals, but the ad-men may well take it over. David Jones

## Evolution of bee dances

SIR — von Frisch<sup>1</sup> and his students are famous for their descriptions of the highly evolved dance 'language' in *Apis* which is used by worker bees throughout this genus to communicate the distance and direction of sources of food, water and nest sites. Because only minor variation in dance biology exists across the genus, hypotheses for the evolution of dance language are problematic. The classical theory<sup>1</sup> suggests that the excited movements of foraging workers became fixed as a stylized round dance. Evolutionary refinements then produced additional dance elements encoding more precise information.

*A. andreniformis* is an Asian species of honey bee only recently recognized as biologically distinct from its sister species, *A. florea*<sup>2</sup>. Like *A. florea*, *A. andreniformis* constructs a nest comprising a single wax comb suspended from a branch. The comb surrounds a section of the supporting branch and the portion of the nest above the branch is used for honey storage. Worker bees cover the comb in a protective curtain. During a study of reproductive isolation of sympatric *A. florea* and *A. andreniformis* through the temporal separation of mating flights<sup>3</sup>, we observed the behaviour of drones before mating flights, recording several hundred flights and many tens of dances using a video camera.

Other than during or somewhat before the time of drone flight, drones of both species are hard to detect among the bees of the protective curtain. During the half hour before flight begins, drones of both species appear on the protective curtain, walk upward and eventually begin flights from the honey storage area at the crown of the nest. Before flying, some drones of *A. andreniformis* run, with wings somewhat extended to the side, in circling loops. Some runs, but not all, end with the drone taking flight. These runs are visually identical in form and tempo to the round dances of *A. mellifera* workers. The dances of drones stimulate other drones in two ways: (1) after encountering a dancing drone, other drones will often follow the dance; and (2) when a dance ends in flight, the following drones often take flight with the lead dancer. This results in groups of two to ten drones taking flight together. Dances and group flights occurred every 2–7 minutes in the three colonies we observed.

We can think of three hypotheses to explain the adaptive value of drone dancing. First, the dance may warm muscles before flight; second, it may help orientate the dancer to celestial cues; or third, it may synchronize group flight by drones. The first two hypoth-

eses lack appeal as most drones do not dance. Mating flights by small groups of *A. andreniformis* may enhance mate location, mating or avoidance of predation. Interestingly, *A. florea* drones also fly in groups but do not dance<sup>3</sup>.

**Thomas E. Rinderer**

**Benjamin P. Oldroyd**

**H. Allen Sylvester**

USDA Agricultural Research Service,  
1157 Ben Hur Road,  
Baton Rouge, Louisiana 70820, USA

**Siriwat Wongsiri**

Department of Biology,  
Chulalongkorn University,  
Bangkok 10330, Thailand

**Lilia I. de Guzman**

Department of Entomology,  
Louisiana State University,  
Baton Rouge, Louisiana 70893, USA

1. von Frisch, K. *The Dance Language and Orientation of Bees* (Harvard University Press, 1967).

2. Wongsiri, S. et al. *Apidologie* **21**, 47–52 (1990).

3. Rinderer, T. E. et al. *J. apic. Res.* (in the press).

## Of mice and men

SIR — Increasing numbers of transgenic mice are now being generated which have heterozygous or homozygous non-functional genes known to be important in humans. There is also a strong expectation in the current literature that mice and men should have essentially the same molecular and cellular characteristics.

An illustration of this is provided in Ed Harlow's recent *News and Views* article (*Nature* **359**, 270–271; 1992). Harlow finds it surprising that mice heterozygous for the Rb tumour-suppressor gene do not develop retinoblastomas, as humans do. Yet in the case of tumour progression, it has been known for many years that mice and men are totally different. Mouse cells can readily be transformed to a tumorigenic phenotype by carcinogens or activated oncogenes, whereas human cells are extraordinarily resistant to such transformation. Moreover, mouse cells first become immortalized (for example, to a 3T3 fibroblast cell line) and then fully transformed. Human cells can be first morphologically transformed by some viral genes, or immortalized, but only very occasionally does a fully transformed immortalized cell line emerge. Therefore, I am not surprised that tumour-suppressor genes and oncogenes behave very differently in mouse and man.

**Robin Holliday**

CSIRO Laboratory for Molecular Biology,  
PO Box 184 North Ryde,  
New South Wales 2113, Australia

## Protein kinase C downregulation?

SIR — de Vries-Smits *et al.*<sup>1</sup> provide novel information on factors that regulate 'downstream' cellular kinases, such as ERK2. Whereas *ras* appears to be important, the reported findings suggest that protein kinase C (PKC) is not involved in insulin-stimulated ERK2 activation. These negative findings on PKC largely derive from an experimental paradigm, 'PKC downregulation' (depletion), provoked by prolonged treatment with certain PKC activators such as the phorbol ester 12-*O*-tetradecanoyl phorbol 13-acetate (TPA).

Unfortunately, this experimental paradigm is commonly flawed. For example, the persistence of insulin effects and the loss of acute TPA effects on ERK2 following 'PKC downregulation' could logically be interpreted to suggest that PKC is not involved during insulin action. But as we have found<sup>2,3</sup>, TPA may be very effective in acutely translocating (activating) and chronically downregulating (depleting) PKC- $\alpha$ , but may have little or no effect on PKC- $\beta$  in certain cell types (for example BC3H-1 myocytes). Moreover, insulin acutely translocates PKC- $\beta$  in both non-downregulated and 'PKC-downregulated' myocytes, but has little or no effect on PKC- $\alpha$  in these cells<sup>3</sup>. These differential effects of insulin and TPA on PKC- $\alpha$  and PKC- $\beta$ , both acutely and chronically, could explain a seemingly paradoxical situation in which insulin-stimulated, PKC-dependent effects persist after prolonged TPA treatment of myocytes, whereas acute TPA-stimulated, PKC-dependent effects are lost.

de Vries-Smits *et al.* used an antibody that specifically recognizes PKC- $\alpha$  (ref. 4) to verify 'PKC downregulation' in their studies in A14 cells, and were careful to point out that the insulin effect on ERK2 is not mediated by a TPA-sensitive PKC. Nevertheless, the unwary reader should be aware of the caveat we describe here in using the experimental paradigm of phorbol ester-stimulated 'PKC downregulation'.

**Robert V. Farese**

**Mary L. Standaert**

**Denise R. Cooper**

J. A. Haley Veterans' Administration  
Hospital,  
University of South Florida College  
of Medicine,  
Tampa, Florida 33612, USA

1. de Vries-Smits, A. et al. *Nature* **357**, 602–604 (1992).

2. Cooper, D. R. et al. *Biochem. biophys. Res. Commun.* **161**, 327–334 (1989).

3. Standaert, M. L., Cooper, D. R., Hernandez, H., Arnold, T. & Farese, R. V. *Endocrinology* (in the press).

4. Walker, J. M., Homan, E. C. & Sando, J. J. *J. biol. Chem.* **265**, 8016–8021 (1990).



# Unmixing of hot inorganic melts

SIR — Phase separation with decreasing temperature occurs in silicate, phosphate and borate liquids and glasses. Phase separation with increasing temperature (a lower critical solution temperature, LCST) is known in various organic systems but has neither been proposed nor documented for oxide liquids. I suggest that an LCST is possible in oxides, present thermodynamic arguments based on the breakdown of complex species with increasing temperature, and summarize circumstantial evidence consistent with an LCST.

The figure (a) shows immiscibility with decreasing temperature in a silicate. The stabilizing contribution of the entropy of mixing ( $T\Delta S$ ), which increases with increasing temperature, overpowers the destabilizing energetics (positive  $\Delta H$ ), causing the miscibility gap to close. This is not the only topology for phase separation. An LCST is known in various organic molecular and polymeric systems (b in the figure). A closed loop of immiscibility (c) or an 'hourglass' shape (d) have been seen.

Immiscibility which increases with increasing temperature can be explained on thermodynamic grounds. In a system A–B, the energy of interaction between A and B is positive (destabilizing) if A and B differ significantly in size and bonding requirements. However, if A and B form a complex or bond strongly, their interaction is negative (stabilizing). Such complexes, though energetically favourable, represent a melt structuring which lowers its entropy. Therefore complexes tend to dissociate on heating. If A–B interactions are unfavourable, immiscibility may set in as the concentrations of free A and B increase. If A–B interactions are only moderately unfavourable, this miscibility gap may close again at high temperature.

Can a similar mechanism apply to silicates, borates and phosphates? Phase-equilibrium studies have not reported an LCST. However, it may not be easy to observe an LCST. Because phase separation occurs at high temperature, where both diffusion and speciation changes are rapid, the phase-separated regime might be difficult to quench and observe. Once a homogeneous liquid is detected above the liquidus, few people have continued heating that liquid several hundred degrees higher to see if it

unmixes. If an LCST is not considered, people may attribute seemingly inconsistent observations to experimental problems, or simply not publish them. Several investigators have brought to my attention unpublished data for borates and phosphates, which do not fit the expected picture of immiscibility and which could suggest an LCST. The most concrete example occurs in  $\text{La}_2\text{O}_3\text{--B}_2\text{O}_3$ . Working with large pots of glass melt, Malmendier<sup>1</sup> noted that lanthanum-rich glasses could be formed only by slow

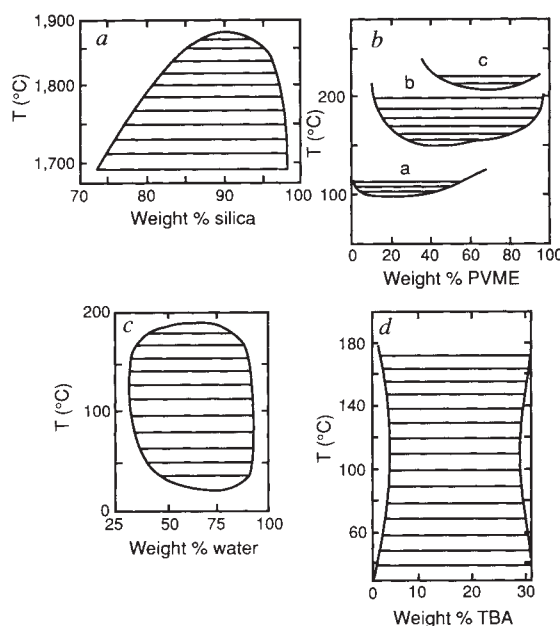
temperature are similar, but the melts in which one mole of  $\text{SiO}_2$  is replaced by  $\text{TiO}_2$  show a much larger heat capacity in the first two hundred degrees above the glass transition, not seen for the titanium-free systems. This may reflect change in Ti coordination and/or dissociation of potassium titanate complexes.

Non-arrhenian viscosity in oxide glasses and melts<sup>2</sup> is consistent with significant change in structure with temperature. Liquids classified as 'fragile' by Angell<sup>4</sup> represent systems whose structures change sharply with increasing temperature to states of higher energy and higher entropy, consistent with the breakdown of complexes. Negative heats of mixing in (Na, K) aluminosilicate glasses<sup>5</sup>, which appear to be absent in the melts<sup>6</sup>, also suggest temperature-induced dissociation. Thus the stage is set for temperature-induced unmixing if the dissociated species interact unfavourably.

An LCST in oxide melts would have significant implications for ceramic processing and magma evolution. Modern *in situ* high-temperature spectroscopy and careful phase studies of systems well above their liquidus temperatures could test this hypothesis. Candidate ceramic systems might be borates, phosphates or silicates containing rare earths or alkalis and Nb, Mo, Zr or Ti, as well as fluorides and fluorozirconates. Geological melts should be relatively depolymerized and contain significant concentrations of alkalis and large polyvalent cations. Komatiitic liquids meet

the first but not the second criterion. Pressure and water content may affect LCST behaviour in magmas. All these predictions are necessarily tentative: an LCST may show up quite unexpectedly.

**Alexandra Navrotsky**  
Department of Geological and  
Geophysical Sciences,  
Princeton University,  
Princeton,  
New Jersey 08544, USA



Some experimentally observed topologies for immiscibility. a, UCST in the system  $\text{CaO--SiO}_2$  (ref. 7); b, LCST in polystyrene (PS)-polyvinyl methylether (PVME) (ref. 8); a, styrene monomer; b, relative molecular mass 110,000 monodisperse polystyrene (PS); c,  $M_r$  200,000 PS; c, both UCST and LCST in 2,4-dimethylpyridine-water<sup>9</sup>; d, hourglass solvus in  $M_r$  7190 polyethylene glycol-*t*-butyl acetate (TBA)<sup>10</sup>.

cooling, rapid cooling resulting in phase separation and/or crystallization. Malmendier did not investigate this in detail, but suggested that crystallization was influenced by temperature-induced changes in boron coordination. An LCST, with a two-phase liquid at higher temperature quenching to two glasses, one of which crystallizes rapidly, is another plausible explanation.

Complexes, especially between alkalis and large polyvalent cations, have been inferred from phase equilibria, thermochemistry, vibrational spectroscopy and NMR<sup>2</sup>. Can they dissociate with increasing temperature to cause an LCST? The evidence is indirect. A change in speciation is expected to cause a stronger than normal change in physical properties with temperature. Heat capacity measurements for  $\text{K}_2\text{O} \cdot 2\text{SiO}_2$  compared to  $\text{K}_2\text{O} \cdot \text{TiO}_2 \cdot \text{SiO}_2$  and  $\text{K}_2\text{O} \cdot 3\text{SiO}_2$  compared to  $\text{K}_2\text{O} \cdot \text{TiO}_2 \cdot 2\text{SiO}_2$  show such an anomaly<sup>3</sup>. The heat capacities of the titanate and silicate liquids at very high

1. Malmendier, W. J. thesis (Penn State Univ., 1969).
2. Scarfe, C. M. (ed) *Short Course in Silicate Melts* (Min. Ass. Canada, Toronto, 1986).
3. Lange, R. A. & Navrotsky, A. *Eos* **71**, 1650 (1990).
4. Angell, C. A. in *Relaxation in Complex Systems* (eds Ngai, K. & Wright, G. B.) 1–10 (U.S. Dept. Commerce, Springfield, 1985).
5. Hervig, R. L. & Navrotsky, A. *Geochim. cosmochim. Acta* **48**, 513–522 (1984).
6. Rammensee, W. & Fraser, D. G. *Geochim. cosmochim. Acta* **46**, 2269–2279 (1982).
7. Tewhey, J. D. & Hess, P. C. *Phys. chem. Glasses* **20**, 41–53 (1979).
8. Nishi, T. & Kwei, T. K. *Polymer* **18**, 285–290 (1975).
9. Onok, H. A. J. *Phase Theory — The Thermodynamics of Heterogeneous Equilibria* (Elsevier, Amsterdam, 1981).
10. Saeki, S., Kuwahara, N., Nakata, M. & Kaneko, M. *Polymer* **17**, 685–689 (1976).

# The geometric phase

Jeeva Anandan

**When a quantum system evolves so that it returns to its initial physical state, it acquires a 'memory' of this motion in the form of a geometric phase in the wavefunction. This phase has observable consequences in a wide range of physical systems, and its presence has now been convincingly demonstrated, for example, in optical and nuclear magnetic resonance experiments.**

A THEORY of nature, once well understood, may not be expected to have surprises in store for us. It is therefore a testament to the richness of quantum theory that it continues to surprise us many decades after its initial formulation. One such surprise was the discovery in 1959 by Aharonov and Bohm<sup>1</sup> that the behaviour of an electrically charged quantum system can be altered in an apparently non-local manner—that is, without any forces acting on it. If a beam of electrons is split into two and the two beams are made to interfere by passing around a cylinder (Fig. 1), interference fringes are observed in the form of a variation in the intensity of the combined beam. Aharonov and Bohm predicted that if a magnetic field is introduced inside the cylinder, these interference fringes will shift even though no electric or magnetic field lines intersect the beams (so that no forces act on them).

Another surprise was the discovery by Berry<sup>2</sup> that a slowly evolving (adiabatic) quantum system retains a 'memory' of its motion when returned to its original physical state. This 'memory' is termed Berry's phase or the geometric phase. The appearance of a geometric phase has been generalized to the case of non-adiabatic evolution<sup>3</sup> of quantum systems, which helps to underline the purely geometrical nature of the phenomenon (which is to say, it depends purely on the geometry of the pathway along which the system evolves). There have now been several experimental observations of the adiabatic and non-adiabatic geometric phases, and it has become clear that it is related to a wide range of physical phenomena, both in quantum systems (for which the Aharonov–Bohm effect represents one expression) and in classical systems (where the corresponding effect is exemplified by the motion of the Foucault pendulum). Here I shall describe how the geometric phase arises in quantum systems; this is seen most easily by first considering its classical analogue. I shall also review briefly some of the ways in which the geometric phase has been observed experimentally—in optics, neutron optics and nuclear magnetic resonance studies.

## The role of phase

According to quantum theory, the state of a particle at any given time is described by a wavefunction  $\psi$ , which is a complex-valued function of space. As the system evolves in time this wavefunction changes accordingly. In the evolution considered by Berry<sup>2</sup>, the final wavefunction  $\psi'$  is related to the initial wavefunction  $\psi$  by  $\psi' = e^{i\phi}\psi$ , where  $\phi$  is a real number and  $i^2 = -1$ . This means that the observable characteristics of the initial and final states are the same and therefore according to classical physics we cannot distinguish between them. But in quantum mechanics,  $\phi$ , which is the additional phase acquired by the wavefunction, constitutes a 'memory' of the evolution undergone by the system. Berry showed that, in addition to a contribution from the effect of the environment at each instant of time,  $\phi$  has a part that is geometrical in nature.

As a simple example, consider a neutron rotating in a magnetic field. A neutron has spin angular momentum  $S = \frac{1}{2}\hbar$ , where  $\hbar = h/2\pi$  and  $h$  is Planck's constant. With any given state of the spin we can associate a unique direction which may be visualized as the average direction of the spin. We may therefore

define a spin vector  $S$  pointing in this direction with magnitude  $S$ . If the neutron is rotated about this spin direction by an angle  $\theta$ , its wavefunction acquires a phase  $\phi = -\frac{1}{2}\theta$ . For example, in a magnetic field  $B$ , the interaction between  $B$  and the neutron's magnetic moment  $\mu$  (which lies along the spin direction and points in the opposite direction) would rotate the neutron continuously about  $B$ ; this is called the Larmor precession. Therefore if  $B$  is in the direction of the neutron spin, the phase of  $S$  will change continuously.

Suppose now that the direction of  $B$  is changed slowly so as to return eventually to its original direction. The neutron spin direction will be pinned to  $B$  and will move with it slowly (that is, adiabatically) (Fig. 2a). But the total rotation of the neutron about its spin direction is not, in general, equal to just the accumulated rotation owing to its Larmor precession at every instant: there is an additional rotation  $\alpha$ , which results in an additional change of phase of the wavefunction  $\beta = -\frac{1}{2}\alpha$ . This  $\beta$  is Berry's phase.

Berry's approach was to use a set of parameters that define the environment of the quantum system. In the above example, the magnetic field is the environment and these parameters are the components of the magnetic field. The additional rotation  $\alpha$  is equal to the solid angle subtended at the origin by the curve  $\Gamma$  traced out by  $B$  in this three-dimensional parameter space. This result has been verified experimentally; I shall derive it below and show that it is purely geometric in origin.

Berry's phase was reconsidered by Aharonov and Anandan<sup>3,4</sup>, who shifted the emphasis from changes in the environment to the motion of the quantum system. In the above example, for instance, they associated the geometric phase with the motion of the neutron spin and not the motion of the magnetic field. This is useful because there are an infinite number of ways of varying the magnetic field to obtain a given motion of the spin, and Aharonov and Anandan showed that, for all of them, the

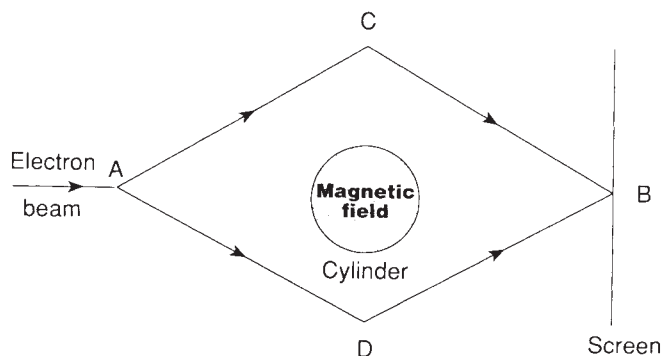


FIG. 1 Schematic illustration of the Aharonov–Bohm effect. An electron beam is split into two at A and the two resulting beams are diverted at C and D so that they interfere in the region surrounding B; the interference pattern is observed on the screen. A magnetic field confined to the cylinder shifts the phase difference between the interfering beams. This results in a shift in the interference pattern, even though the beams move in a field-free region with no forces acting on them.



same geometric phase is obtained which is uniquely associated with the motion of the neutron. This approach enabled them to generalize Berry's phase to non-adiabatic motions. For example, when the magnetic field is changed rapidly (regardless of whether it returns to its original direction), if the neutron spin returns to its original state then it will have undergone an additional rotation equal to the solid angle  $\alpha$  subtended by the curve  $C$  traced out by the spin vector  $S$  (Fig. 2b). This gives rise to a geometric phase  $\beta = -\frac{1}{2}\alpha$ ; but now it is not necessary for the motion to be adiabatic. Note that the neutron also has another state with spin vector  $-S$  which, under the same circumstances, changes by a phase that includes a geometric phase  $\beta = +\frac{1}{2}\alpha$ .

The approach of Aharonov and Anandan is in the spirit of the work of Pancharatnam<sup>5</sup>, who introduced a geometric phase in optics about three decades before Berry, in his study of the rotation of the polarization of light owing to a sequence of measurements. The polarization state of a photon is analogous to the spin of the neutron. It turns out that the Aharonov-Bohm effect and the geometric phase can both be explained by basically the same geometric concept, called parallel transport.

### Classical parallel transport

For simplicity, consider first a vector on a two-dimensional plane which is parallel-transported along a curve: that is, moved along the curve so that its length and direction are constant. Parallel transport may be generalized to a vector moved along a curve on a curved surface such as a sphere by requiring that the vector moves without a change in its magnitude and without rotation about the instantaneous normal to the surface; that is, the component of the angular velocity of the vector in the direction of the normal is zero. This condition may be restated in terms of concepts that are intrinsic to the surface, by noting that a small portion of a smooth surface is like a plane. Therefore parallel transport may be defined along any smooth curve by requiring that, for each infinitesimal segment of the curve, it is the same as parallel transport on a plane: the given vector, the infinitesimal segment of the curve and the parallel-transported vector form three sides of a parallelogram. This restatement enables us to ignore the three-dimensional space in which the surface is embedded. (If the curve is in a curved space of arbitrary dimension, instead of on a two-dimensional surface, parallel transport may be defined in the same way.) In general, a suitable rule or prescription for parallel transport of a vector along any curve in an arbitrary space is called a connection. The special case of the infinitesimal parallelogram rule above is called a Riemannian connection.

Of great importance is the transformation undergone by vectors when they are parallel-transported around a closed curve, which is called a holonomy transformation. If a vector is parallel-transported around a closed curve on the plane it comes back unchanged, but if parallel-transported around a closed curve  $C$  on a sphere it gets rotated by an angle equal to the solid angle subtended by  $C$  at the sphere's centre, which is the holonomy transformation in this case (Fig. 3a). For a sphere of radius  $r$ , this solid angle is given by

$$\alpha = A/r^2 \quad (1)$$

where  $A$  is the area enclosed by the closed curve, and the curvature is  $1/r^2$  everywhere on the sphere. This result can be seen quite easily for the case of a vector, initially at the north pole, which is parallel-transported first to the equator by moving it down the longitude along which the vector points, then a quarter of the way around the equator, and finally back to the north pole along the new longitude. The vector ends up pointing along the new longitude (Fig. 3b). Therefore, even though the vector has not been rotated at any point along its route, nevertheless it arrives back at the north pole rotated by the angle between the two longitudes, namely  $\pi/2$  radians. This is also the solid

angle subtended at the centre of the sphere by the vector's path, as obtained from equation (1).

Parallel transport on a sphere can be illustrated by means of the Foucault pendulum, which is used to measure the rate of rotation of the Earth. Imagine a set (triad) of cartesian axes on the surface of the Earth (approximated as a sphere) with its  $z$  axis always vertical and the  $x$  and  $y$  axes always horizontal. A Foucault pendulum, suspended from a point on the  $z$  axis, oscillates in a vertical plane. If the triad is moved so that its  $x$  and  $y$  axes are parallel-transported along a curve  $C$  on the Earth's surface, the triad acquires no rotational velocity about the  $z$  axis (normal to the surface of the Earth) and therefore the plane of oscillation of the pendulum with respect to this triad does not change. Conversely, we can use the condition that the plane of oscillation of the Foucault pendulum is constant to ensure that the  $x$  and  $y$  axes are being parallel-transported. When the triad returns to its original position, the plane of oscillation has rotated by an angle  $\alpha$ , equal to the solid angle subtended by  $C$  at the centre of the Earth<sup>8</sup>. This is an example of Hannay's angle<sup>9</sup> which is a classical analogue of Berry's phase.

This experiment has so far been performed keeping the point of oscillation fixed with respect to the Earth, so that as the Earth rotates,  $C$  runs around a latitude of the Earth. Then  $\alpha = 2\pi(1 - \cos \theta)$  or  $2\pi \cos \theta$ , depending on the sense in which it is measured (Fig. 3c). But clearly, by moving the point of oscillation appropriately,  $C$  can be any closed curve on the Earth's surface. Thus the Foucault pendulum, which is usually used to demonstrate that the Earth rotates, can just as well be used to demonstrate, in principle, that the Earth's surface is curved. If the Earth were a disk then  $\alpha$  would always be 0 (or  $2\pi$ ), corresponding to parallel transport on a plane. The Earth's radius  $r$  may be determined from equation (1) if we know  $\alpha$  and the area of the Earth's surface enclosed by  $C$ .

In practice such a determination could be made by having a Foucault pendulum on each of two nearby ships with their planes of oscillations parallel. The ships now separate for a while and then come together again so that the closed curve formed by their paths on an imaginary non-rotating sphere that coincides with the Earth's surface enclosed an area  $A$ . The planes of oscillation will no longer be parallel, but will diverge at an

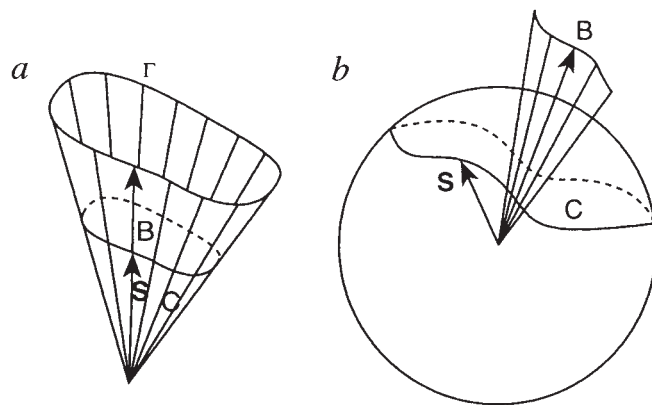


FIG. 2 Geometric phase for a neutron in a magnetic field. *a*, The field  $B$  is varied slowly along the curve  $\Gamma$ . The neutron spin, represented by the vector  $S$ , follows  $B$ . The phase acquired by the neutron is the sum of the dynamical phase that it acquires through its Larmor precession in  $B$ , and the geometric phase, equal in magnitude to half the solid angle  $\alpha$  subtended by  $\Gamma$  at the origin. *b*,  $B$  moves in such a way that the neutron spin vector  $S$  traces the circuit  $C$  on a sphere with radius  $S$ . Unlike case *a*, the magnetic field need not be parallel to  $S$  or vary slowly. The phase acquired is the sum of the dynamical phase resulting from the Larmor precession around the component of  $B$  in the direction of  $S$  at every instant of time, and the geometric phase whose magnitude is half the solid angle subtended by  $C$  at the origin. *a*, is obtained as a special case corresponding to a particular choice of  $B$  that is parallel to  $S$  and slowly varying.

angle  $\alpha$ , and  $r$  is then determined from equation (1). This approach is fundamentally different from that of the ancient Greeks, who measured  $r$  by comparing the shadows of vertical poles at two different points on the Earth's surface, because the latter is a measurement of extrinsic curvature—the curvature of the Earth in the three-dimensional space in which it is embedded—whereas the former is a measurement of intrinsic curvature, defined by intrinsic properties of the surface without any reference to the euclidean geometry of the embedding space.

But the Earth is strictly not a sphere: it is slightly oblate. For an arbitrary surface, the intrinsic curvature  $R$  (sometimes called the gaussian curvature) at any point may be defined as follows: the angle by which a vector rotates when it is parallel-transported around an infinitesimal closed curve around this point is

$$d\alpha = R dA$$

where  $dA$  is the area of the infinitesimal portion enclosed by the curve. For a curve that encloses a finite area, the area can be divided into infinitesimal areas and the angle of rotation  $\alpha$  of a vector parallel-transported around this curve (the holonomy transformation) is the sum of the contributions from the infinitesimal areas:

$$\alpha = \int R dA \quad (2)$$

In general  $R$  can vary from point to point, but for a sphere  $R$  is equal to  $1/r^2$  everywhere. Therefore equation (2) gives equation (1) in this special case.

### Geometric phases and parallel transport

Classical parallel transport should enable us at least to gain some intuitive understanding of the Aharonov-Bohm effect and Berry's phase. The former is analogous to parallel transport on a cone. A cone may be formed by taking a flat piece of paper bounded by two straight edges at an angle  $\theta$  and joining these edges (Fig. 4). As the paper is not stretched or compressed during this process, the cone has no intrinsic curvature except at the vertex, which may be smoothed out so that the curvature is finite everywhere. Thus, except for this region at the vertex, the intrinsic geometry in any small region of the cone is the same as the flat geometry of the initial sheet of paper. Therefore, away from the vertex, a vector will be parallel-transported just as it would be on the flat sheet of paper. But a vector  $V$  parallel-transported into  $V'$  around a closed curve  $C$  that encloses the vertex undergoes a rotation by the angle  $\theta$ . This is the holonomy transformation associated with  $C$ .

The curvature at the vertex can be regarded as analogous to the magnetic field within the cylinder in the Aharonov-Bohm experiment, while the zero intrinsic curvature everywhere else corresponds to the vanishing of the magnetic field outside the cylinder. Then the rotation by the angle  $\theta$  which relates  $V$  and  $V'$  (Fig. 4) is analogous to the phase difference between the two interfering beams, which they acquire by travelling in the field-free region, just as  $V$  moves in the curvature-free region. The fact that the Aharonov-Bohm effect cannot be understood in terms of forces suggests that this analogy with parallel transport should be taken seriously.

This analogy is valid more generally when both electric and magnetic fields are present. This suggests that the electromagnetic field may be a connection for parallel transport of the value of the wavefunction  $\psi(x, t)$  along a path in space-time, and the Aharonov-Bohm effect arises from the corresponding holonomy transformation around the closed curve followed by the interfering beams. The statement that the electromagnetic field is such a connection is the same as saying that it is a gauge field. The electric and magnetic field strengths at each point in space-time constitute the curvature of this connection at that point, with the connection represented by the electromagnetic potential.

Classical parallel transport can also help us to understand

Berry's phase for the neutron in a magnetic field, mentioned previously (Fig. 2a), as pointed out by Anandan and Stodolsky<sup>6</sup>. As the direction of the magnetic field  $B$  changes slowly, the spin vector  $S$  (which follows  $B$ ) will trace out a curve  $C$  on a sphere with radius  $S$ . Imagine now a cartesian triad moved around  $C$  with its  $z$ -axis radial and its  $x$ - and  $y$ -axes being parallel-transported. It returns rotated about the  $z$  axis by the solid angle  $\alpha$  given by equation (1). Similarly, a neutron spin vector that maintains its orientation relative to the triad will also be rotated by the same angle  $\alpha$ . This rotation, the origin of which is purely geometrical, changes the phase of the spin wavefunction by  $\beta = -\frac{1}{2}\alpha$  because the axis of rotation here is along the direction of the spin vector. At each instant, the neutron also undergoes a Larmor precession with frequency  $\omega_L = 2\mu B/\hbar$ , where  $\mu$  is the magnitude of its magnetic moment. The accumulation of these rotations leads to an additional phase change, called the dynamical phase, owing to the dynamical interaction between  $\mu$  and  $B$ .

Even this simple example shows the magnitude of Berry's achievement: he effectively identified a new way of rotating the neutron spin which had been missed by all the physicists who for many years had studied the properties of the neutron. This rotation is purely geometrical, so that, unlike the Larmor precession, it is not necessary to know the values of the dynamical quantities such as  $\mu$ ,  $B$  or even  $\hbar$  in order to determine it<sup>6</sup>.

But by focusing on the motion of  $S$  instead of  $B$  we can see that the two restrictions on  $B$  imposed by Berry—adiabaticity and a need for  $B$  to be parallel or antiparallel to  $S$ —may be removed. Consider the more general case in which  $S$  traces out the same curve  $C$  but  $B$  need be neither slowly varying nor in the direction of the neutron spin (Fig. 2b). There is an infinite number of ways of changing  $B(t)$  as a function of time  $t$  to make this happen. At any instant  $t$  there is a Larmor precession of the neutron with frequency  $\omega_L$  about the direction of  $B(t)$ . Hence the angular velocity of the neutron about its spin direction is equal to the projection of the total angular velocity onto the spin direction. Therefore the angular frequency of the precession about the spin axis with respect to the triad

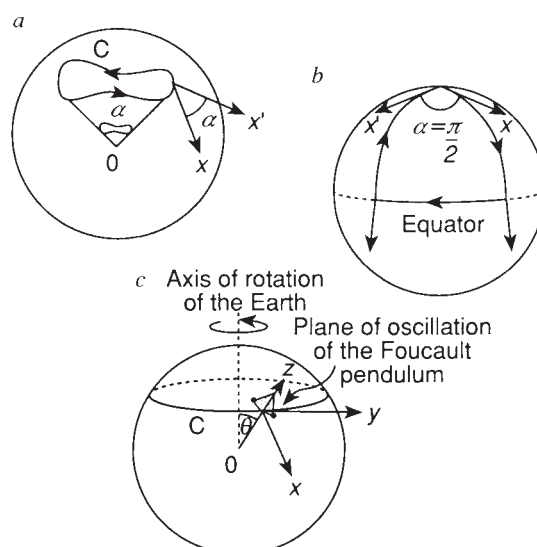


FIG. 3 a, A vector  $X$  when parallel-transported around a closed curve  $C$  on a sphere returns as  $X'$ , which is related to  $X$  by a rotation by an angle equal to the solid angle subtended by  $C$  at the centre of the sphere. b, Verification of this result for the special case of parallel transport down a longitude, around the equator by  $\pi/2$ , and back up the new longitude. In this case  $\alpha = \pi/2$ . c, Parallel transport illustrated by the Foucault pendulum. The imaginary cartesian triad moves along the latitude  $C$  so that its  $z$ -axis is radial while the  $x$ - and the  $y$ -axis are parallel-transported on the Earth's surface. The plane of oscillation of the pendulum is constant with respect to this triad. But it comes back rotated by the same angle  $\alpha$  by which the triad has rotated about the  $z$ -axis because of parallel transport around  $C$ .



parallel-transported along  $C$  is  $\omega_L \cos \gamma = (2\mu B \cos \gamma)/\hbar$ , where  $\gamma$  is the angle between  $\mathbf{B}$  and  $\mathbf{S}$ . The accumulated rotation owing to this precession gives the dynamical phase. This generalizes the dynamical phase of Berry, which corresponds to the special case  $\gamma = 0$ . But as we have seen, the triad parallel-transported around  $C$  comes back rotated by the solid angle  $\alpha$ , so that the neutron's spin wavefunction acquires an additional, geometric phase  $\beta = -\frac{1}{2}\alpha$ . The geometric phase depends only on the motion of the system, specified by the curve  $C$ , and is independent of the orientation or rate of change of the magnetic field that causes this motion.

## Quantum parallel transport

Thus we can understand the geometric phase for spin by considering classical parallel transport on a sphere and its effect on the wavefunction. For more-complex quantum systems this argument can be generalized using group theory, in which case  $\alpha$  is generalized to a set of geometric angles<sup>6,7</sup>, described briefly later. But the origin of the geometric phase in an arbitrary quantum system can also be understood by considering parallel transport within the context of quantum theory, in which the state of a system is described by a vector.

There are two important fundamental differences between classical and quantum physics. First, in quantum physics one may form a new state of a system by adding two initial states (which is why quantum states are represented by vectors), whereas in classical physics adding dynamical states is not possible. More generally, we can form any linear combination of two quantum states (see Box 1). For example, for electron beams interfering around a cylinder (Fig. 1) the intensity distribution in the interference region can be explained by supposing that the state of each electron in this region is the sum of two states corresponding to the two beams, as if the electron traveled simultaneously along both beams. But according to classical physics, the electron can travel along only one or other of the beams, giving rise to a different intensity distribution that does not accord with experiment.

Another important difference is that, if a quantum system is known to be in a state described by the vector  $|\psi_1\rangle$  and a measurement is performed to see if it is in a state with vector  $|\psi_2\rangle$ , quantum theory predicts a probability for the success of this experiment, called the probability of transition. In classical physics, on the other hand, a measurement made on a system in a given state will not find it in another state. The quantum probability of transition is evaluated from the 'probability amplitude' for a transition from state  $|\psi_1\rangle$  to  $|\psi_2\rangle$ , a complex number that depends only on the two state vectors. The probability amplitude is written  $\langle\psi_2|\psi_1\rangle$ , and is also called the inner product between  $|\psi_1\rangle$  and  $|\psi_2\rangle$  (see Box 1). We can write  $\langle\psi_1|\psi_2\rangle = A e^{i\phi}$ , where  $\phi$  is a real number representing the phase difference between  $|\psi_1\rangle$  and  $|\psi_2\rangle$ , and  $A$  is a non-negative real number not exceeding unity. Then  $|\langle\psi_1|\psi_2\rangle|^2 = A^2$  is the probability of transition from  $|\psi_1\rangle$  to  $|\psi_2\rangle$ .

Thus  $A$  corresponds to an observable quantity. But does  $\phi$ , which may seem to arise from a mathematical formalism, also have observable physical consequences? We shall see that the geometric phase is a direct physical consequence of this phase difference. This aspect of the probability amplitude was perhaps first realized by Pancharatnam<sup>5</sup>, who showed that if two state vectors  $|\psi_1\rangle$  and  $|\psi_2\rangle$  have the same phase (that is,  $\phi = 0$ ), and  $|\psi_2\rangle$  and  $|\psi_3\rangle$  also have the same phase, then  $|\psi_1\rangle$  and  $|\psi_3\rangle$  need not have the same phase. The non-transitivity of the relation of two state vectors having the same phase can be understood in terms of parallel transport, and it is this non-transitivity that is ultimately responsible for the geometric phase<sup>4,12-14</sup>.

To determine the physical state of a quantum system, we need to measure a set of compatible observables. If the state vector  $|\psi\rangle$  gives a certain set of observed values for these observables, then  $|\psi'\rangle = c|\psi\rangle$ , where  $c$  is any complex number, will give the same values (see box). The set of all vectors obtained by multi-

plying a given vector  $|\psi\rangle$  by all possible complex numbers is called a ray. It follows that all vectors in a given ray represent the same physical state because they give the same values for the measured observables.

The set of rays of the Hilbert space  $\mathcal{H}$  (see Box 1) is called the projective Hilbert space<sup>3</sup>, denoted by  $\mathcal{P}$ . For neutron spin states, for example,  $\mathcal{P}$  can be identified with the set of spin vectors  $\mathbf{S}$  which form the sphere that I used earlier to explain the geometric phase for the neutron. Consider a curve  $p$  in  $\mathcal{P}$ , such that  $p(s)$  denotes the point on this curve corresponding to the parameter value  $s$ . Also, for each value of  $s$  let  $|\psi(s)\rangle$  be a vector in the ray corresponding to  $p(s)$  such that  $|\psi(s)\rangle$  varies smoothly along the curve  $p(s)$ . Two natural conditions for parallel-transporting  $|\psi(s)\rangle$  along a curve are that the length of  $|\psi(s)\rangle$  is preserved (that is,  $\langle\psi(s)|\psi(s)\rangle$  is independent of  $s$ ) and that  $|\psi(s)\rangle$  and  $|\psi(s+ds)\rangle$  have the same phase (that is,  $\langle\psi(s)|\psi(s+ds)\rangle$  is real and positive). These two conditions imply that

$$\langle\psi(s)|\frac{d}{ds}|\psi(s)\rangle = 0 \quad (3)$$

which was obtained in a different manner by Simon<sup>15</sup>.

Equation (3) is the rule for parallel transport of  $|\psi(s)\rangle$  along any curve in  $\mathcal{P}$ , and therefore defines a connection over  $\mathcal{P}$ . It indicates that the change in  $|\psi(s)\rangle$  during parallel transport,  $d|\psi(s)\rangle/ds$ , has a zero inner product with  $|\psi(s)\rangle$ , so that the change in  $|\psi(s)\rangle$  cannot be along its own direction. As  $\langle\psi(s)|\psi(s)\rangle$  is independent of  $s$ , the only way  $|\psi(s)\rangle$  can change along its own direction is by being multiplied by a phase factor  $e^{i\epsilon}$ , which may be regarded as a 'rotation' of  $|\psi(s)\rangle$ . Hence, forbidding such a 'rotation' is analogous to the no-rotation condition in the definition of parallel transport of a vector on a sphere described earlier. If, however,  $|\psi(s)\rangle$  is parallel-transported around a closed curve ( $p(\tau) = p(0)$  for some  $\tau$ ), it comes back 'rotated':

$$|\psi(\tau)\rangle = e^{i\theta} |\psi(0)\rangle \quad (4)$$

### BOX 1 Some fundamental principles of quantum mechanics

According to the notation introduced by Dirac, a state (vector) of a quantum system is denoted by the 'ket'  $|\psi\rangle$ . Given two states  $|\psi_1\rangle$  and  $|\psi_2\rangle$ , any linear combination  $c_1|\psi_1\rangle + c_2|\psi_2\rangle$ , where  $c_1$  and  $c_2$  are complex numbers, represents a possible new state. If  $\mathcal{H}$  denotes the set of possible states of a quantum system then the above property is contained in the statement that  $\mathcal{H}$  forms a vector space over the field of complex numbers. Given two states  $|\psi\rangle$  and  $e^{i\theta}|\psi\rangle$ , where  $\theta$  is real and  $i^2 = -1$ , we say that  $\theta$  is the phase difference between the two states.

The inner product of two given states  $|\psi_1\rangle$  and  $|\psi_2\rangle$  is a complex number, denoted  $\langle\psi_1|\psi_2\rangle$ . The vector space  $\mathcal{H}$  with a suitable inner product is called a Hilbert space. An observable is a measurable property of a quantum system and is represented by a hermitian operator that acts on the Hilbert space  $\mathcal{H}$ . During a measurement of an observable, the possible values obtained, which are real, are the eigenvalues of the corresponding hermitian operator. (The eigenvalues of an hermitian operator are always real.)

The dynamical evolution of a state is given by the Schrödinger equation.

$$i\hbar \frac{d}{dt} |\psi\rangle = H |\psi\rangle$$

where  $H$  is a linear hermitian operator acting on the states in  $\mathcal{H}$ , called the hamiltonian, and  $\hbar = h/2\pi$ ,  $h$  being Planck's constant. It follows that the Schrödinger equation is linear in the sense that, if  $|\psi_1(t)\rangle$  and  $|\psi_2(t)\rangle$  are two solutions of this equation, the complex linear combination  $c_1|\psi_1(t)\rangle + c_2|\psi_2(t)\rangle$ , where  $c_1$  and  $c_2$  are arbitrary complex numbers, is also a solution.

The wavefunction of a quantum state is defined by  $\psi(\mathbf{r}, t) = \langle\mathbf{r}|\psi(t)\rangle$ , where  $|\mathbf{r}\rangle$  is the state corresponding to the system being in position  $\mathbf{r}$ . Then  $|\psi(\mathbf{r}, t)|^2$  is the probability per unit volume of finding the system at position  $\mathbf{r}$  at time  $t$ .

Thus the factor  $e^{i\beta}$  arises from the holonomy transformation in this case, and  $\beta$  is the geometric phase. This additional phase is acquired because of the curvature of the connection on  $\mathcal{P}$ . I will now derive an explicit expression for  $\beta$ .

### Calculating the geometric phase

We say that an evolution  $|\psi(t)\rangle$  is cyclic in the time interval  $t=0$  to  $\tau$  if  $|\psi(\tau)\rangle = e^{i\phi}|\psi(0)\rangle$  for some phase  $\phi$ . As points in  $\mathcal{P}$  have no phase information, the beginning and end points of this evolution must correspond to the same point in  $\mathcal{P}$ , so that the system moves around a closed curve in  $\mathcal{P}$ . We now define a basis vector field  $|\tilde{\psi}(t)\rangle$  such that  $|\psi(t)\rangle = e^{i\varphi(t)}|\tilde{\psi}(t)\rangle$  and  $|\tilde{\psi}(\tau)\rangle = |\tilde{\psi}(0)\rangle$ . In this way we express the phase of  $|\psi(t)\rangle$  in terms of  $|\tilde{\psi}(t)\rangle$ . This is analogous to describing the vector on a curved surface by its components with respect to a chosen coordinate system on the surface. The parallel transport of the vector can be described in terms of these components. Part of the phase change  $\phi$  is due to the quantum parallel transport of  $|\psi(t)\rangle$ , which can be expressed in terms  $|\tilde{\psi}(t)\rangle$ . This phase, denoted  $\beta$ , generalizes to an arbitrary quantum system the geometric phase obtained earlier for the neutron. It can be shown<sup>3</sup>, when  $|\psi(t)\rangle$  evolves according to the Schrödinger equation (see box), that  $\phi = \beta + \delta$ , where

$$\delta = -\hbar^{-1} \int_0^\tau \langle \psi(t) | H(t) | \psi(t) \rangle dt \quad (5)$$

is the dynamical phase and

$$\beta = i \int_0^\tau \langle \tilde{\psi}(t) | \frac{d}{dt} | \tilde{\psi}(t) \rangle dt = i \oint_C \langle \tilde{\psi} | d | \tilde{\psi} \rangle \quad (6)$$

is the geometric phase; here  $d$  is the differential operator on  $\mathcal{P}$  and  $H(t)$  is the Hamiltonian operator, the eigenvalues of which are the energy levels of the system.

The phase  $\beta$  is geometrical because of the following reasons: (1) Unlike  $\delta$ ,  $\beta$  is independent of the rate at which the system travels along the unparametrized curve  $C$ . (2) Unlike  $\delta$ ,  $\beta$  is independent of the particular hamiltonian  $H(t)$  producing motion along  $C$ . There are, in fact, an infinite number of hamiltonians that cause a given motion of the quantum systems along  $C$ , but the same geometric phase  $\beta$  is obtained for all of them. (3) The geometric phase factor  $e^{i\beta}$  is independent<sup>4</sup> of the chosen  $|\psi\rangle$ . Hence the factor  $e^{i\beta}$  depends only on  $C$ —it corresponds to the holonomy transformation associated with  $C$  for the connection defined by equation (3). In other words, if  $|\psi\rangle$  is parallel-transported around  $C$ , it gets multiplied by  $e^{i\beta}$ . Statements (1), (2) and (3) show that  $\beta$  is geometrical in origin because it depends only on the motion of the quantum system defined by  $C$  and not on the particular interaction undergone by the system, which is determined by  $H$ .

For the simple case of the neutron considered earlier, equations (5) and (6) yield

$$\delta = -\frac{1}{2} \int_0^\tau \omega_L(t) \cos \gamma(t) dt, \quad \beta = -\frac{1}{2} \alpha \quad (7)$$

That is,  $\delta$  is the accumulated phase change owing to the component of the angular velocity of the Larmor precession in the spin direction, and  $\beta$  is acquired by the additional rotation owing to parallel transport, as explained earlier. More generally, equation (5) gives the accumulated phase change owing to the instantaneous change in  $|\psi(t)\rangle$  in the direction of  $|\psi(t)\rangle$ , according to Schrödinger's equation, while  $\beta$  is the additional phase due to parallel transport. But the parallel-transported state vector undergoes the holonomy transformation, giving  $|\psi\rangle$  an additional phase according to equation (6). It has been shown that<sup>2,3</sup>, in a sense, the Aharonov-Bohm effect is a manifestation of this phase.

**Adiabatic approximation.** An important special case of the geometric phase results from the adiabatic cyclic evolution

studied by Berry<sup>2</sup>. Suppose that a hamiltonian  $H$  is a function of a set of parameters  $\mathbf{R}$ . By varying  $\mathbf{R}(t)$  sufficiently slowly with time  $t$ ,  $H(\mathbf{R})$  can be made to vary adiabatically. The adiabatic theorem in quantum mechanics states that if a quantum state is initially an eigenstate of  $H(\mathbf{R}(0))$ , it continues to remain as the eigenstate of the 'snapshot' hamiltonian  $H(\mathbf{R}(t))$ :

$$H(\mathbf{R}(t))|\psi(t)\rangle = E(t)|\psi(t)\rangle \quad (8)$$

In particular, if  $\mathbf{R}$  moves around a closed curve so that  $\mathbf{R}(0) = \mathbf{R}(\tau)$ , the initial and final states are eigenstates of  $H(\mathbf{R}(0))$ . If  $H(\mathbf{R}(t))$  is non-degenerate (that is, the eigenstate is unique up to a phase factor), then  $|\psi(\tau)\rangle = e^{i\phi}|\psi(0)\rangle$ . The evolution is then cyclic, with  $\phi = \beta + \delta$ , and from equations (5) and (8),  $\delta = -\hbar^{-1} \int_0^\tau E(t') dt'$  in this special case.

Even though adiabaticity is not essential for a geometric phase to be acquired, there are important examples of adiabatic evolution in physics. It occurs, for example, in molecular physics, which forms the basis of all of chemistry. The electronic states of a vibrating or rotating molecule can be determined by calculating the states for stationary nuclei and allowing these to evolve adiabatically as the nuclei move; the coordinates of the electrons and nuclei are called the fast and slow variables respectively. This is the Born-Oppenheimer approximation. As the nuclei travel around a closed path in the parameter space of slow variables, the electronic wavefunction acquires a geometric phase<sup>16-19</sup>.

### Geometric angles

I will now generalize the treatment of the geometric phase for the neutron to a particle with arbitrary angular momentum, such as an atom or a molecule. A particle with an angular momentum characterized by the quantum number  $J$  (an integer or half-integer) has  $2J+1$  independent states, which undergo a change of phase when the particle is rotated about some chosen axis, called the quantization axis. These states are denoted by  $|n\rangle$ ,  $n = J, J-1, \dots, -J+1, -J$ , where the rotation of the state  $|n\rangle$  by an angle  $\theta$  about this axis results in a phase change of  $-n\theta$ . For the neutron  $J = \frac{1}{2}$  and the two states  $n = \frac{1}{2}, -\frac{1}{2}$  correspond to the spin vectors  $\mathbf{S}$  and  $-\mathbf{S}$  parallel and antiparallel to the quantization axis.

Suppose that the particle is in a magnetic field  $\mathbf{B}(t)$  which varies arbitrarily in some interval of time  $t=0$  to  $\tau$ . The effect of this on the spin states is to give all of them the same rotation at every instant of time. Therefore the quantization axis also undergoes a rotation. The net effect induced by the magnetic field during the time interval  $t=0$  to  $\tau$  is a rotation about a unique axis, which must therefore come back to itself. Hence each of the spin states  $|n\rangle$  with this axis chosen as the quantization axis undergoes a cyclic evolution, changing only by a phase factor  $\phi_n$ . The possible directions of the quantization axis may be represented by points on a sphere. So the effect of  $\mathbf{B}(t)$  in this time interval is to move the quantization axis along a closed curve  $C$  on this sphere.

The instantaneous precession of each state  $|n\rangle$  about the instantaneous quantization axis is the component of the Larmor

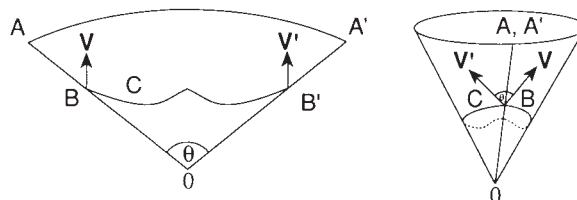


FIG. 4 Parallel transport on a cone as a classical analogy of the Aharonov-Bohm effect. The cone is formed by aligning  $OA$  and  $OA'$ . The vector  $\mathbf{V}$  parallel-transported from  $B$  around the curve  $C$  on the cone would come back to  $B$  (identified with  $B'$ ) as  $\mathbf{V}'$ , rotated by the angle  $\theta$ . This is analogous to the Aharonov-Bohm phase shift between beams interfering around  $C$  which encloses a magnetic field; the field corresponds to the curvature at the apex of the cone.



frequency in this direction. These accumulate to yield the dynamical phase  $\delta_n$ , given by equation (7) with  $n$  replacing the factor of  $\frac{1}{2}$ . But as we saw by parallel-transporting a triad around  $C$ , each state should come back rotated by the additional angle  $\alpha$ , the solid angle subtended by  $C$ . Therefore it acquires the additional geometric phase  $\beta_n = \phi_n - \delta_n = -n\alpha$ . This  $\beta_n$  can also be obtained directly from equation (6).

The angle  $\alpha$  is purely geometrical and is independent of the quantum numbers  $J$  or  $n$ , unlike the geometric phase which is proportional to  $n$ . For a more general physical system, however, the evolution of the system cannot be understood as an ordinary rotation. Instead,  $\alpha$  generalizes to a set of angles  $\alpha_1, \alpha_2, \dots, \alpha_m$ , for quantum systems<sup>6,7</sup> as well as classical systems<sup>7,9</sup>. The classical case can be obtained as the limiting case of the states  $|n\rangle$  being so-called WKB (this stands for Wentzel, Kramers and Brillouin) states, which have fairly well defined momenta or actions and therefore poorly defined positions because of the Heisenberg uncertainty relation. These states can be associated with ensembles of classical particles with definite momentum or action. In this limit, the angles  $\alpha_k$  are the same as Hannay's angles<sup>9,10</sup> and their non-adiabatic generalization<sup>7,11</sup>, which determine the geometric part of the cyclic evolution undergone by a classical system. Interesting physical applications of these angles in classical physics have been described by Hannay<sup>9</sup> and Berry and Hannay<sup>11</sup>.

### Geometric phase from measurements

According to the 'Copenhagen interpretation' of quantum mechanics, a quantum state undergoes two types of change. Between measurements it undergoes a continuous evolution governed by the Schrödinger equation; but when a measurement is made, it undergoes a discontinuous change into an eigenstate of the observable being measured. Pancharatnam's pioneering work<sup>5</sup> amounted to identifying a geometric phase in the evolution of a system of photons, caused by the second type of change.

Consider a beam of polarized light, which consists of photons in a given spin state, passing through a sequence of polarizers such that the final polarizer restores the initial polarization. The photons then undergo a cyclic evolution (Fig. 5). Each polarizer acts as a 'measuring device' and the polarization of a photon that passes through it undergoes a discontinuous change. For a given momentum, a photon may be regarded as having two possible independent polarization states, which define a two-dimensional Hilbert space analogous to the spin state of the neutron (although the photon has spin 1). Therefore, the projec-

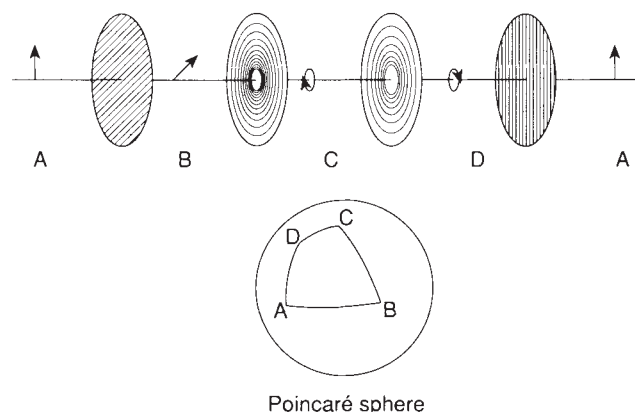


FIG. 5 Pancharatnam's optical experiment which first introduced the geometric phase<sup>5</sup>. Polarized light passes through a sequence of polarizers each of which lets through only the component of light with a particular polarization. A photon therefore goes through a sequence of polarization states represented by the points A, B, C, ..., A on the Poincaré sphere. The final polarizer restores the initial polarization; but a net phase is acquired by the photon equal in magnitude to half the solid angle of the geodesic polygon ABC, ..., A on the sphere.

tive Hilbert space is again a sphere, called the Poincaré sphere, points on the surface of which represent the polarization states (without defining their phase). The successive polarization states of the photon correspond to a sequence of points on this sphere.

When a photon passes through each polarizer, the photon's state vectors immediately before and after the passage have the same phase as defined earlier, which was shown by Pancharatnam<sup>5</sup>. Berry<sup>12</sup> showed that this is equivalent to parallel-transporting, with respect to the connection defined by equation (3), the photon polarization state vector along the shortest geodesic joining the two points on the Poincaré sphere corresponding to the initial and final states. Therefore for a cyclic evolution the state vector undergoes the holonomy transformation obtained by parallel-transporting it around the geodesic polygon formed from the sequence of points on the Poincaré sphere; and so the photon state acquires a phase of magnitude  $\alpha/2$ , where  $\alpha$  is the solid angle of the geodesic polygon—in agreement with Pancharatnam's result<sup>5</sup>. This result has been generalized to cyclic evolution of an arbitrary quantum system induced by measurements, in which case the geometric phase is obtained by parallel-transporting around the geodesic polygon defined by the sequence of states of the system in the corresponding projective Hilbert space<sup>4,14</sup>.

### Experimental tests

The acquisition of a geometric phase by a spin has been verified experimentally by Bitter and Dubbers<sup>20</sup> for the special case of adiabatic evolution. They passed a beam of spin-polarized neutrons through a solenoid such that, in the rest frame of the neutron, the magnetic field  $\mathbf{B}$  slowly swept out a cone of solid angle  $\alpha$ . The neutron spin states parallel and antiparallel to  $\mathbf{B}$  acquire, in addition to their dynamical phases, the Berry phases  $\pm\frac{1}{2}\alpha$ . This should result in an additional precession of the neutron by the angle  $\alpha$  (the difference between the Berry phases), as Bitter and Dubbers observed. A similar experiment was performed by Suter *et al.*<sup>21</sup>, who used techniques of nuclear magnetic resonance (NMR) spectroscopy to subject a spin to an effective magnetic field which slowly sweeps out a cone.

An ingenious experiment was performed by Tycko<sup>22</sup>, who used a crystal of sodium chlorate in which the spins of the chlorine nuclei were aligned with the crystal's symmetry axis, which is here the quantization axis. He excited these nuclei into a superposition of two spin states by using a pulse of radio waves. By rotating the crystal about some other axis, the two spins states then acquired geometric phases that increased with time, producing a difference in the frequency (change of phase per unit time) of the two states. This shift was detected by a change in the frequency of the radiation emitted by the nuclei, measured using standard NMR phase-sensitive detection techniques. The use of mechanical rotation in this experiment makes it clear that the geometric phase is associated with the motion of the system and is independent of the particular hamiltonian inducing the motion<sup>3,4</sup>.

The non-adiabatic geometric phase was also observed by Suter *et al.*<sup>23</sup> using NMR techniques. They studied a system of coupled protons, a quantum system with total spin  $J = 1$ . As seen previously, there are three independent states so that the Hilbert space  $\mathcal{H}$  is three-dimensional. A two-level subspace was made to undergo a cyclic evolution in  $\mathcal{H}$  by applying a time-dependent magnetic field. The geometric phase was measured by looking at interference between these two states and the third, unperturbed level, and the prediction of Aharonov and Anandan<sup>3</sup> was confirmed for different circuits in  $\mathcal{H}$ .

An interesting optical experiment was performed by Chiao and colleagues<sup>24,25</sup>, who sent polarized light through a helically twisted optical fibre so that the direction of light on entry to and exit from the fibre was the same. As the light passes through the fibre its direction traces out a closed curve  $C$  over the sphere of all possible directions. On exit, the plane of polarization is

rotated relative to its initial state by the solid angle  $\alpha$  subtended by C at the centre of the sphere. This is because the photons, for polarized light, are in a superposition of spin states parallel and antiparallel to their momentum which acquire geometric phases  $\mp\alpha$ . This corresponds to a rotation of the photon spin (polarization) by the angle  $\alpha$ , because the photon has spin 1.

This experiment can also be understood as a special case of the general treatment of Anandan and Stodolsky<sup>6</sup>, described earlier. Consider a cartesian triad with its  $z$  axis in the direction of the momentum (along the axis of the fibre), which is moved along this axis without rotation about it. With respect to this triad, the plane of polarization does not change<sup>3,26-28</sup>. But the triad can be regarded as undergoing parallel transport along C and therefore comes back rotated by  $\alpha$  about its  $z$  axis, and the plane of polarization is rotated likewise. A quantum-mechanical description is not needed to understand this experiment because the geometry is so general as to apply to both classical and quantum systems. Again the geometric phase arises entirely from the motion of the system, with adiabaticity being irrelevant.

For further discussion of the diverse ramifications and

experimental applications of geometric phases see refs 29 and 30.

## Conclusion

The purpose of physics is to simplify and unify our understanding of nature. Geometric phases and angles provide a unified description of a wide range of phenomena in classical and quantum physics. But their greatest value lies perhaps in giving us a new way of looking at quantum theory. It is useful to look at a theory from different viewpoints, because it may be easier to see how to modify the theory from one perspective than from another. Historically, this has been particularly true of geometric reformulations of physical theories. It may not be unreasonable to hope, therefore, that the new insights gained in the past few years through the study of the geometric phase and related geometric structures in quantum mechanics may have heuristic value. □

Jeeva Anandan is in the Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208, USA. This work was performed partially at the Department of Theoretical Physics, University of Oxford, Oxford OX1 3NP, UK.

- Aharonov, Y. & Bohm, D. *Phys. Rev.* **115**, 485-491 (1959).
- Berry, M. V. *Proc. R. Soc. A* **392**, 45-57 (1984).
- Aharonov, Y. & Anandan, J. *Phys. Rev. Lett.* **58**, 1593-1596 (1987).
- Anandan, J. & Aharonov, Y. *Phys. Rev. D* **38**, 1863-1870 (1988).
- Pancharatnam, S. *Proc. Indian Acad. Sci. A* **44**, 247-262 (1956); reprinted in *Collected Works of S. Pancharatnam* (Oxford Univ. Press, 1975).
- Anandan, J. & Stodolsky, L. *Phys. Rev. D* **35**, 2597-2600 (1987).
- Anandan, J. *Phys. Lett. A* **129**, 201-207 (1988).
- Kugler, M. & Shtrikman, S. *Phys. Rev. D* **37**, 934-937 (1988).
- Hannay, J. H. *J. Phys. A* **18**, 221-230 (1985).
- Berry, M. V. *J. Phys. A* **18**, 15-27 (1985).
- Berry, M. V. & Hannay, J. H. *J. Phys. A* **21**, L325-L331 (1988).
- Berry, M. V. *J. Mod. Opt.* **34**, 1401-1407 (1987).
- Ramaseshan, S. & Nityananda, R. *Current Science* **55**, 1225-1226 (1986).
- Samuel, J. & Bhandari, R. *Phys. Rev. Lett.* **60**, 2339-2342 (1988).
- Simon, B. *Phys. Rev. Lett.* **51**, 2167-2170 (1983).
- Herzberg, G. & Longuet-Higgins, H. C. *Disc. Farad. Soc.* **35**, 77-82 (1963).
- Stone, A. J. *Proc. R. Soc. Lond. A* **351**, 141-150 (1976).
- Mead, C. A. & Truhlar, D. G. *J. chem. Phys.* **70**, 2284-2296 (1979).
- Moody, J., Shapere, A. & Wilczek, F. in *Geometric Phases in Physics* (eds Shapere, A. & Wilczek, F.) 160-183 (World Scientific, Singapore, 1989).
- Bitter, T. & Dubbers, D. *Phys. Rev. Lett.* **59**, 251-254 (1987).
- Suter, D., Chingas, G. R., Harris, Pines, A. *Molec. Phys.* **61**, 1327-1340 (1987).
- Tycko, R. *Phys. Rev. Lett.* **58**, 2281-2284 (1987).
- Suter, D., Mueller, K. T. & Pines, A. *Phys. Rev. Lett.* **60**, 1218-1220 (1988).
- Chiao, R. Y. & Wu, Y. *Phys. Rev. Lett.* **57**, 933-936 (1986).
- Tomita, A. & Chiao, R. Y. *Phys. Rev. Lett.* **57**, 937-940 (1986).
- Berry, M. V. *Nature* **326**, 277-278 (1987).
- Rytov, S. M. *Dokl. Akad. Nauk. SSSR* **XVIII**, 263 (1938); reproduced in *Topological Phases in Quantum Theory* (eds Markovski, B. & Vinitzky, S. I.) (World Scientific, Singapore, 1989).
- Haldane, F. D. M. *Phys. Rev. Lett.* **59**, 1788 (1987).
- Geometric Phases in Physics* (eds Shapere, A. & Wilczek, F.) (World Scientific, Singapore, 1989).
- Zwanziger, J. W., Koenig, M. & Pines, A. *Rev. Phys. Chem.* **41**, 601-646 (1990).

## ARTICLES

# $\alpha$ -Inhibin is a tumour-suppressor gene with gonadal specificity in mice

Martin M. Matzuk<sup>\*†</sup>, Milton J. Finegold<sup>†</sup>, Jyan-Gwo J. Su<sup>‡</sup>, Aaron J. W. Hsueh<sup>‡</sup> & Allan Bradley<sup>\*</sup>

<sup>\*</sup> Institute for Molecular Genetics and <sup>†</sup> Department of Pathology, Baylor College of Medicine, Houston, Texas 77030, USA

<sup>‡</sup> Division of Reproductive Biology, Department of Gynecology/Obstetrics, Stanford University, Stanford, California 94305, USA

**The inhibins are  $\alpha:\beta$  heterodimeric growth factors that are members of the transforming growth factor- $\beta$  family. To understand the physiological roles of the inhibins in mammalian development and reproduction, a targeted deletion of the  $\alpha$ -inhibin gene was generated by homologous recombination in mouse embryonic stem cells. Mice homozygous for the null allele (inhibin-deficient) initially develop normally but every mouse ultimately develops mixed or incompletely differentiated gonadal stromal tumours either unilaterally or bilaterally. Inhibin is thus a critical negative regulator of gonadal stromal cell proliferation and the first secreted protein identified to have tumour-suppressor activity.**

THE inhibins and activins are developmentally and physiologically important dimeric growth factors<sup>1-4</sup> with structural homology to a large group of proteins including the transforming growth factor ( $\text{TGF}$ )- $\beta$ s (ref. 5), Mullerian-inhibiting substance (MIS)<sup>6</sup>, and multiple proteins related to the *Xenopus* Vg1 protein and the *Drosophila* DPP-C protein<sup>7</sup>. Similar to other members of the  $\text{TGF}$ - $\beta$  family, these proteins are secreted from cells as dimers and interact with cell surface receptors<sup>8</sup>. The inhibins are  $\alpha:\beta$  heterodimers, whereas the activins are  $\beta:\beta$  dimers.

There are two unique but homologous  $\beta$ -subunits,  $\beta\text{A}$  and  $\beta\text{B}$ , which are shared between these growth factors (Fig. 1a)<sup>1</sup>.

The inhibins and activins were initially discovered as gonadal peptides that either inhibited or stimulated pituitary follicle stimulating hormone (FSH) production, respectively<sup>1</sup>. Several studies suggest, however, that the inhibins and activins may have critical intragonadal paracrine and/or autocrine roles, often acting as mutual antagonists. The major gonadal sites of inhibin synthesis are the Sertoli cells in males and the granulosa



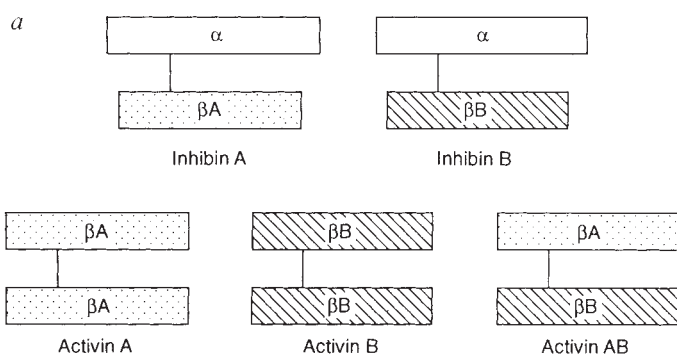


FIG. 1 Structures of the inhibins and activins and comparison of the disulphide dimer forms present in wild-type and homozygote inhibin-deficient mice. **a**, There are two inhibin heterodimers, inhibin A ( $\alpha:\beta A$ ) and inhibin B ( $\alpha:\beta B$ ). These forms share common  $\beta$ -subunits with the activin  $\beta:\beta$  dimers. There are three activins, activin A ( $\beta A:\beta A$ ), activin B ( $\beta B:\beta B$ ) and activin AB ( $\beta A:\beta B$ ), which differ in combination of  $\beta$ -subunits. **b**, Wild-type mice contain the 3 activin (A, B or AB) and 2 inhibin (A or B) dimers described

**b**

|                               | Activin |   |    | Inhibin |   |
|-------------------------------|---------|---|----|---------|---|
|                               | A       | B | AB | A       | B |
| Wild-type                     | +       | + | +  | +       | + |
| $\Delta\alpha/\Delta\alpha$   | +       | + | +  | -       | - |
| $\Delta\beta A/\Delta\beta A$ | -       | + | -  | -       | + |
| $\Delta\beta B/\Delta\beta B$ | +       | - | -  | +       | - |

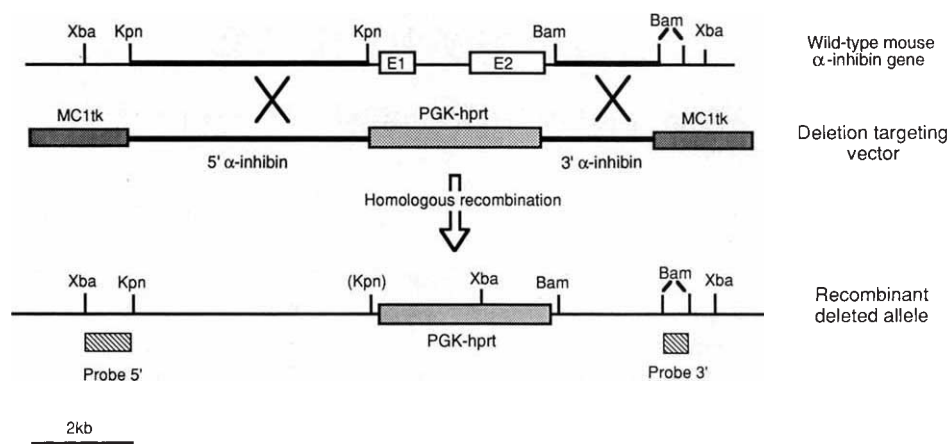
above. Mice homozygote for the targeted deletion of the  $\alpha$ -subunit gene ( $\Delta\alpha/\Delta\alpha$ ) lack the  $\alpha$ -subunit and therefore cannot make the two inhibin forms. Because  $\beta A$  and  $\beta B$  subunits are still present, the three activin forms can still be synthesized. This contrasts to hypothetical mice lacking the  $\beta A$  subunit gene ( $\Delta\beta A/\Delta\beta A$ ) or  $\beta B$  subunit gene ( $\Delta\beta B/\Delta\beta B$ ), which would be deficient in both activin and inhibin forms. Minus, absent; plus, present.

cells in females<sup>1</sup>. In the male, inhibin inhibits spermatogonia DNA synthesis<sup>9</sup> and reduces spermatogonia number<sup>10</sup>, whereas activin stimulates spermatogonial proliferation *in vitro*<sup>11</sup>. Inhibin synthesis by the Sertoli cells fluctuates during the stages of spermatogenesis<sup>12</sup>. The binding of inhibin to different populations of germ cells<sup>13</sup> also changes during the various stages of spermatogenesis, supporting the possibility that there is an important paracrine interaction between these intratubular cell types (that is, Sertoli and germ cells) which may be mediated by inhibin. Furthermore, inhibin and luteinizing hormone stimulate steroidogenesis in Leydig cells *in vitro*<sup>14</sup>. In females, inhibin seems to have homologous functions because it has been suggested to inhibit oocyte meiosis<sup>15</sup>, increase the number of ovarian follicles<sup>16</sup> and stimulate thecal cell steroidogenesis *in vitro*<sup>14</sup>. Activin has an antagonistic function compared to inhibin on ovarian follicular development, causing follicular atresia<sup>16</sup>, and recombinant activin A induces granulosa cell proliferation *in*

*vitro*<sup>17</sup>. Other studies have shown that the inhibins and activins cause diverse responses in cell types and tissues other than the pituitary and gonads, and in the majority of cases, the effects of these growth factors are also mutually antagonistic<sup>1</sup>. Analysis of the  $\alpha$ ,  $\beta A$ , and  $\beta B$  messenger RNAs demonstrates that they are expressed in multiple tissues during adult and fetal life<sup>3,4,18,19</sup>. These reports indicate that inhibin is synthesized during embryonic development in the gonads and placenta and during adult life in the gonads, pituitary, adrenal gland, spleen and nervous system. In particular, inhibin and activin have been suggested to be critical regulators of placental human chorionic gonadotropin (hCG) synthesis<sup>20</sup>. Although the inhibins and activins may have important intragonadal and extragonadal roles, it is difficult to interpret the physiological significance in mammals of these *in vitro* and *in vivo* studies because of lack of good experimental models.

Several studies of dimeric proteins related to the inhibins and

FIG. 2 Targeted deletion of the  $\alpha$ -inhibin gene in ES cells. The two exons (E1 and E2) of the mouse  $\alpha$ -inhibin gene are contained on an 11.1-kb *Xba*I fragment as described<sup>26</sup>. The deletion targeting vector (pMM18) contains from left to right, an MC1-tk (thymidine kinase) expression cassette (transcription from left to right), 4.8 kb of 5'  $\alpha$ -inhibin homology (the sequence between the two *Kpn*I sites with both sites destroyed), a PGK-hprt expression cassette (transcription from left to right), 1.2 kb of 3'  $\alpha$ -inhibin homology (the sequence between the two *Bam*H1 sites) and a second MC1-tk cassette (transcription from right to left). The  $\alpha$ -inhibin gene, which was used for vector constructions, was isolated from an ICR mouse genomic library<sup>26</sup> and therefore was non-isogenic compared to the 129Sv/Ev-derived ES cells which were being used for the gene targeting experiment. The backbone for the vector is pBluescript SK+ (Stratagene). The vector was linearized with *Kpn*I, leaving the pSK+ backbone attached to the MC1-tk on the right side. The homologous regions of the endogenous wild-type  $\alpha$ -inhibin allele and the targeting vector are represented as bold lines and the hypothetical crossovers between these segments are shown by the crossed lines. The resultant recombinant allele has deleted 3.4 kb of endogenous sequence (*Kpn*I to *Bam*H1) containing exons 1 and 2 and replaced it with the 3.6-kb PGK-hprt expression cassette. The presence of the PGK-hprt expression cassette also introduces a new *Xba*I restriction site into the locus which is used for detection of the recombinant allele with Southern



blot analysis. The probes used for Southern blot analysis are the 1.0-kb *Xba*I to *Kpn*I fragment (5' probe) and the 0.4-kb *Bam*H1 fragment (3' probe). For clarity of representation, the  $\alpha$ -inhibin exon sizes are not drawn to scale. PGK, Phosphoglycerate kinase promoter; hprt, hypoxanthine phosphoribosyl transferase. *Xba*I, *Xba*I; *Kpn*I, *Kpn*I; *Bam*H1, *Bam*H1; (Kpn), *Kpn*I site destroyed by the recombination event.

**METHODS.** Linearized vector (25  $\mu$ g) was electroporated into  $1 \times 10^7$  AB2.1 ES cells as described<sup>27</sup>. ES cell clones were selected in M15 medium<sup>45</sup> containing HAT (hypoxanthine, aminopterin and thymidine) and FIAU<sup>27</sup> and analysed as described<sup>30</sup>. An average of 243 double-resistant clones per  $1 \times 10^7$  cells electroporated was found, which was a sevenfold enrichment compared to selection in HAT alone.

activins have led to the hypothesis that they may be involved in tumorigenesis as 'anti-proliferative' growth factors. For instance, although the normal function of MIS is to cause regression of the Mullerian ducts in males, it also inhibits the growth of some reproductive tract carcinoma cell lines<sup>6,21</sup>. Similarly, the TGF- $\beta$ s antagonize the proliferative effect of several growth factors and inhibit growth of a number of cell types *in vitro*<sup>5</sup>. By these criteria, these proteins can be hypothesized to fall into the class of tumour-suppressor proteins. Although several tumour-suppressor genes have been identified, and on the basis of the protein structure can be assigned to different positions in the signal transduction cascade<sup>22</sup>, none of these genes encodes an extracellular ligand.

A major tool to determine the function of specific gene products in mammals and to generate animal models for disease states is the use of gene targeting strategies in murine embryonic stem (ES) cells<sup>23</sup>. Although this strategy has been widely used to examine the roles of genes in development, the generation of homozygous viable animals that have a mutation on both chromosomal copies also provides a very sensitive *in vivo* test for tumour-suppressor activity of a mutated gene<sup>24</sup>. Although ES cell/gene targeting technology has provided *in vivo* models which mimic the tumour-suppressor activity of specific genes identified in familial human syndromes<sup>24</sup>, it has not yet identified novel tumour-suppressor genes.

To understand the function of inhibin in mammalian reproduction and development, we have generated mice heterozygous and homozygous for a deletion of the inhibin  $\alpha$ -subunit gene. Because inhibin is an  $\alpha : \beta$  heterodimer whereas activin is a  $\beta : \beta$  dimer, we chose to target the  $\alpha$ -subunit gene because deficiency of the  $\alpha$ -subunit will result in a deficiency of the inhibins without causing a deficiency in the activins (Fig. 1b). Male and female mice heterozygous for the deleted  $\alpha$ -subunit gene are normal and fertile. Male and female mice homozygous for the deleted  $\alpha$ -subunit gene, and therefore inhibin-deficient, are viable. However, these inhibin-deficient mice are susceptible to the development of mixed or incompletely differentiated gonadal stromal tumours which can be detected in some cases as early as 4 weeks of age. These studies implicate inhibin as a critical intragonadal negative regulator of gonadal cell proliferation. The predisposition of inhibin-deficient mice to tumours identifies inhibin as the first extracellular (ligand) protein that can function as a tumour suppressor.

### Targeting of the $\alpha$ -inhibin gene

To target the  $\alpha$ -inhibin gene in murine ES cells, a positive-negative selection strategy<sup>25</sup> was used. The targeting vector (Fig. 2) contained a total of 6 kilobases (kb) of homology flanking the  $\alpha$ -inhibin coding sequences<sup>26</sup>. A PGK-hprt expression cassette was used as a positive selectable marker and two MC1-tk expression cassettes<sup>27</sup> were used for negative selection. Because the targeting vector contained sequences 5' and 3' to the two exons coding for the  $\alpha$ -inhibin gene, correct homologous recombination between the targeting vector and one of the  $\alpha$ -inhibin alleles in the ES cells should result in a deletion of 3.4 kb of the  $\alpha$ -inhibin gene containing the entire coding sequence. This would ensure that no  $\alpha$ -inhibin mRNA and therefore no  $\alpha$ -inhibin protein was produced in animals homozygous for the targeted allele.

The  $\alpha$ -inhibin targeting construct was electroporated into the hprt-negative AB2.1 ES cell line<sup>28</sup> carrying the TG4 mutant *hprt* allele *hprt*<sup>mb2</sup> (ref. 29). HAT- and FIAU-resistant clones (see legend to Fig. 2) were screened by Southern blot analysis<sup>30</sup> using 5' and 3' DNA probes to identify targeted events (Fig. 2). Two ES cell clones,  $\alpha$ 1-B2 and  $\alpha$ 1-H10, out of 450 clones screened showed the predicted fragments generated by homologous recombination. The wild-type  $\alpha$ -inhibin gene is encoded on a 11.1-kb *Xba*I fragment detected both by the 5' and the 3' probes (Fig. 3). Because the PGK-hprt expression cassette introduces a novel *Xba*I site into the target locus, an *Xba*I digest of DNA

TABLE 1 Gonadal pathology in inhibin-deficient mice

| Males<br>Case(s)   | Age<br>(wks) | Histopathological classification                                                                                                                                                                        |
|--------------------|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1                  | 4            | Intratubular, undifferentiated gonadal stromal tumour (unilateral)                                                                                                                                      |
| 2-6                | 5            | Intratubular, focally invasive, mixed and undifferentiated gonadal stromal tumour (bilateral)                                                                                                           |
| 7                  | 5            | Intratubular, focally invasive undifferentiated gonadal stromal tumour (unilateral)                                                                                                                     |
| 8                  | 6            | Multifocal, mixed and focally invasive undifferentiated gonadal stromal tumour (unilateral)                                                                                                             |
| 9                  | 6            | Multifocal, mixed (bilateral) and focally invasive (unilateral) undifferentiated gonadal stromal tumour                                                                                                 |
| 10-12              | 6            | Multifocal, mixed and focally invasive undifferentiated gonadal stromal tumour (bilateral)                                                                                                              |
| 13-46              | 7-17         | Multifocal, mixed, invasive undifferentiated gonadal stromal tumour (bilateral)                                                                                                                         |
| 47                 | 9            | Multifocal, mixed, invasive undifferentiated gonadal stromal tumour (unilateral)                                                                                                                        |
| Females<br>Case(s) | Age<br>(wks) | Histopathological classification                                                                                                                                                                        |
| 1                  | 4            | Central granulosa-thecal cell tumours (bilateral); granulosa cell hyperplasia in follicles (bilateral)                                                                                                  |
| 2                  | 4            | Granulosa cell hyperplasia in follicles (bilateral)                                                                                                                                                     |
| 3                  | 5            | Central gonadal stromal tubular tumours (bilateral); granulosa cell hyperplasia in follicles (bilateral)                                                                                                |
| 4-5                | 5-6          | Granulosa-thecal cell tumour (unilateral); gonadal stromal tubular tumour (unilateral); granulosa cell hyperplasia in follicles (bilateral) and focal, atypical granulosa cell hyperplasia (unilateral) |
| 6                  | 6            | Multifocal, invasive granulosa-thecal cell tumour (unilateral); Sertoli cell-like component; granulosa cell hyperplasia in follicles (bilateral)                                                        |
| 7-8                | 6            | Multifocal, mixed, invasive gonadal stromal tumour (unilateral); granulosa cell hyperplasia (bilateral)                                                                                                 |
| 9                  | 8            | Multifocal, mixed, invasive gonadal stromal tumour (unilateral); granulosa cell hyperplasia (bilateral)                                                                                                 |
| 10-23              | 8-16         | Multifocal, mixed, invasive gonadal stromal tumour (bilateral)                                                                                                                                          |
| 24                 | 9            | Multifocal, mixed, invasive gonadal stromal tumour (unilateral); haemorrhagic cysts (unilateral)                                                                                                        |

isolated from the targeted clones yields a wild-type fragment of 11.1 kb detected both by the 5' and the 3' probes and additional recombinant fragments of 7.2 kb or 4.1 kb detected by the 5' and the 3' probes, respectively. These new fragments correspond to the deleted  $\alpha$ -inhibin allele and are present at equal intensity to the wild-type allele present on the other chromosome. These results confirm that the two ES cell clones contained a deleted  $\alpha$ -inhibin allele (a null allele).

### Generation of $\alpha$ -inhibin-deficient mice

The ES cell clone  $\alpha$ 1-B2 was used to generate chimaeric mice by blastocyst microinjection, and both male and female offspring heterozygous for the deleted  $\alpha$ -inhibin allele were detected among the agouti progeny of the chimaeras, proving that animals heterozygous for a deleted  $\alpha$ -inhibin allele are viable.

Because the major site of inhibin synthesis is in the gonads, the effect on fertility of having only one intact  $\alpha$ -inhibin allele was determined by crossing heterozygote males with C57BL/6 (wild-type) females. From these crosses, litters were generated of normal size and frequency, and half of the offspring were heterozygote for the targeted allele, with an equal representation of males and females. Thus, there appears to be neither gene-distortion nor sex-distortion in the generation of heterozygote mice.



To determine the role of  $\alpha$ -inhibin in development, homozygous inhibin-deficient animals were generated by crossing heterozygote mice. Litter sizes were normal and 670 offspring have been genotyped. Of these, 167 were homozygotes (25%), 323 were heterozygotes (48%) and 180 were wild-type (27%) animals, consistent with the expected mendelian frequency of 1:2:1. Southern blot analysis (Fig. 3*d, e*) of a representative litter at 3 weeks of age revealed that two of the offspring were homozygous for the deleted  $\alpha$ -inhibin allele and therefore inhibin-deficient. Furthermore, out of the 167 homozygotes, 90 were male and 77 were female. Thus, inhibin-deficient male and female mice survive embryonic life, are viable, and appear to have normal overt sexual differentiation.

To confirm that the homozygote mice were devoid of the  $\alpha$ -subunit coding sequences, the Southern blot in Fig. 3*d* was stripped and rehybridized with an  $\alpha$ -inhibin complementary DNA probe (Fig. 3*e*). As expected, the cDNA probe detected an 11.1-kb fragment in the two animals genotyped as wild-type in Fig. 3*d*, but failed to hybridize to any fragments in the two animals genotyped as homozygotes in which the recombinant allele was designed to lack the two coding exons. Because the

$\alpha$ -inhibin gene was deleted, no  $\alpha$ -inhibin mRNA or protein could be synthesized.

### Inhibin-deficient mice develop tumours

The homozygous animals were initially healthy and had normal external genitalia; however, the mice tested in the initial studies proved to be infertile in crosses with wild-type mice. To determine the cause of this infertility, the gonads were examined (Table 1). Essentially every homozygote animal analysed (all 47 males and 23 out of 24 females) demonstrated macro- or microscopic evidence of gonadal tumours. Detailed histological analysis identified these tumours as mixed or incompletely differentiated gonadal stromal tumours<sup>31-33</sup>. The tumours were evident as early as 4 weeks of age in both sexes as microscopic foci of nodular proliferation and appeared to be clonal (Fig. 4*a-d*). In the males, testicular enlargement and grossly visible foci of haemorrhage (Fig. 4*g*) were usually evident by 5 weeks of age. The tumours, often of a mixed phenotype, were focally invasive (Fig. 4*h*). Although spermatogenesis was initially active in these males from 5-7 weeks, regression occurs in parallel with the enlargement of the tumour mass. The number of Leydig

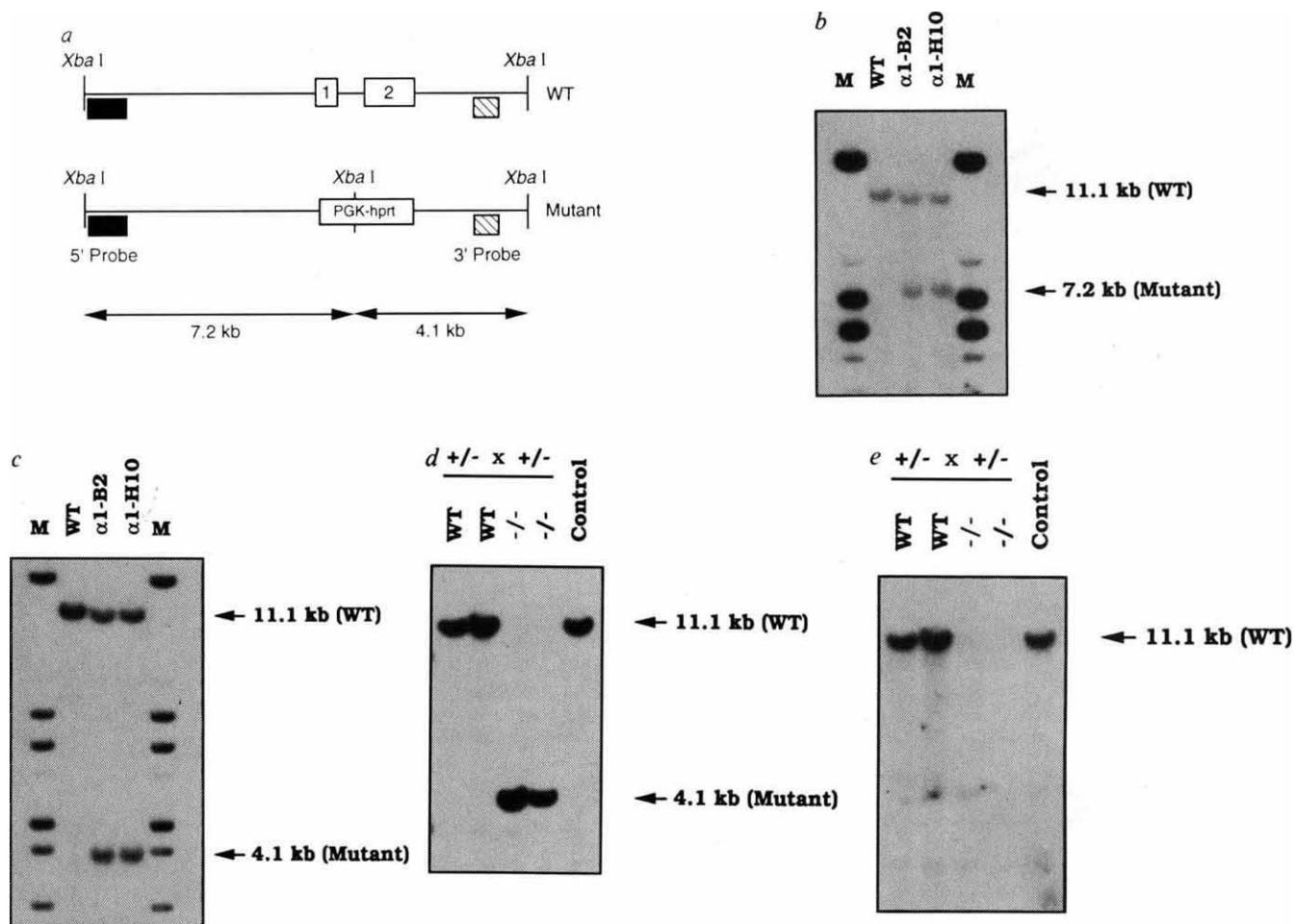


FIG. 3 Southern blot analysis of ES cell DNA and offspring from heterozygous matings. ES cell DNA (5  $\mu$ g) from the wild-type (WT) AB2.1 cell line and from two putative positive clones ( $\alpha$ 1-B2;  $\alpha$ 1-H10) were restricted with *Xba*I, electrophoresed on a 0.7% agarose gel, blotted onto Gene Screen Plus nylon membranes (Dupont-NEN) and hybridized with either a 5'  $\alpha$ -inhibin gene probe (b) or a 3'  $\alpha$ -inhibin gene probe (c) as shown in a. The WT 11.1-kb *Xba*I fragment (indicated) is detected in the 3 ES cell DNA lanes with either 5' or 3' probes. A 7.2-kb fragment is detected with the 5' probe in the  $\alpha$ 1-B2 and  $\alpha$ 1-H10 lanes but not in the WT lane. Similarly, a 4.1-kb fragment is detected with the 3' probe in the lanes with the two clones but not in the WT lane. These two fragments of equal intensity to the WT allele

confirm the presence of a deleted  $\alpha$ -inhibin allele in these two clonal lines. *d, e*, Genomic DNA (5  $\mu$ g), isolated from the tails of 21-day-old pups generated from a cross of two heterozygote mice, was restricted with *Xba*I, subjected to Southern blot analysis and probed with either a 3'  $\alpha$ -gene probe (*d*) or a human  $\alpha$ -inhibin cDNA probe (*e*). The control lane contains genomic DNA isolated from a wild-type animal. M,  $\lambda$  BstEII DNA markers hybridized with a  $\lambda$ -specific probe.

METHODS. Blastocyst microinjection, reimplantation, and the generation of these chimaeric males with a high (50-95%) contribution of the ES cell-derived (129Sv/Ev strain) agouti pigmentation to the coat colour were performed as described<sup>27,45</sup>.

cells also decrease from 5–7 weeks compared to age-matched controls. Analysis of the ovaries from 7–16-week-old females revealed analogous haemorrhage and disruption of the normal follicular architecture by the invasive gonadal stromal tumours. The ovarian tumours that develop in these 7–16-week-old females are often mixed, and can be described as either nodular masses of granulosa and thecal cells, undifferentiated gonadal stromal derivatives, or clusters of mitotically active stromal cells in a tubule arrangement reminiscent of seminiferous tubules (Fig. 4e, f).

The appearance of these tumours in the inhibin-deficient mice contrasts with the absence of such tumours in either wild-type animals<sup>24</sup> (up to 11 months of age) or heterozygote mice (up to 9 months). The absence of early tumour formation in the males and females heterozygous for the deleted  $\alpha$ -inhibin gene further assures us that the insertion of the expression cassette into the  $\alpha$ -inhibin locus does not cause a dominant phenotype in these mice. Furthermore, testis descent is normal and thus the development of tumours in the inhibin-deficient males cannot be secondary to testis maldescent, a known cause of human testicular cancer. The development of normal secondary sexual organs and mature-appearing gametes in both male and female homozygotes indicates apparently normal hormonal activities during embryogenesis and adulthood. Thus, the reason for infertility in these inhibin-deficient mice is not yet understood but is probably secondary to the early development of the tumours.

Because inhibin is believed to be a key regulator of pituitary FSH production and FSH levels have been postulated to be involved in gonadal cell proliferation, radioimmunoassays of serum FSH were performed on serum obtained from inhibin-deficient mice and age-matched controls (Table 2). Comparison of the serum FSH concentrations in older male and female inhibin-deficient mice (with overt tumours present) revealed a two- to threefold increase in FSH levels compared to heterozygote or wild-type controls. A similar increase in FSH levels were observed in the adolescent inhibin-deficient males compared to controls and in the adolescent females compared to the wild-type controls but not compared to the heterozygotes.

## Discussion

To understand the physiological roles of the inhibins in mammalian development and reproduction, a targeted deletion of the  $\alpha$ -inhibin gene (a null allele) was generated by homologous recombination in murine embryonic stem cells. Using a positive-negative selection strategy<sup>25</sup>, two ES cell clones out of 450 HAT-resistant, FIAU-resistant clones screened contained the  $\alpha$ -inhibin null allele which was transmitted through the mouse germ line. Intercrosses of mice heterozygous for the  $\alpha$ -inhibin null allele generated the expected 25% of offspring homozygous for the  $\alpha$ -inhibin null allele (that is, inhibin deficient). Both male and female inhibin-deficient mice initially appeared healthy with no apparent gross abnormalities.

In this study, we chose to delete the  $\alpha$ -inhibin subunit gene rather than one of the two  $\beta$ -subunit genes. This is a critical point in the experimental design because the deletion of one of the  $\beta$ -subunits would delete both activin and inhibin activity (Fig. 1b). The phenotype of an animal lacking the  $\beta$ -subunit could not be conclusively established as a consequence of lack of the activities of activin, inhibin or both. However, the absence of the  $\alpha$ -subunit gene in our experiments abolishes both inhibin A and B but does not result in a deficiency in activin (Fig. 1b).

Earlier studies by Roberts *et al.*<sup>19</sup> had shown that the  $\alpha$ -inhibin mRNA seemed to be present only in the gonads of male and female mice beginning at ~14 days post-coitum. However, the gonads and external genitalia of male and female inhibin-deficient mice were normal before tumour development by gross examination and detailed histology. Thus, expression of inhibin in the embryonic gonad is not essential for normal sexual differentiation and development. Previous studies had also shown that  $\alpha$ -inhibin mRNA is present in rat and human

TABLE 2 Serum FSH concentrations

| Mice assayed       | Mean $\pm$ s.d.   | n  |
|--------------------|-------------------|----|
| Males 7–12 weeks   |                   |    |
| Homozygote         | 277.8 $\pm$ 144.4 | 10 |
| Heterozygote       | 93.7 $\pm$ 33.4   | 10 |
| Wild type          | 122.4 $\pm$ 45.2  | 10 |
| Males 4–5 weeks    |                   |    |
| Homozygote         | 128.2 $\pm$ 99.2  | 5  |
| Heterozygote       | 41.5 $\pm$ 9.9    | 5  |
| Wild type          | 54.2 $\pm$ 10.9   | 4  |
| Females 7–13 weeks |                   |    |
| Homozygote         | 231.6 $\pm$ 125.4 | 5  |
| Heterozygote       | 74.3 $\pm$ 20.7   | 5  |
| Wild type          | 103.4 $\pm$ 82.9  | 5  |
| Females 3–5 weeks  |                   |    |
| Homozygote         | 84.4 $\pm$ 31.6   | 5  |
| Heterozygote       | 109.3 $\pm$ 75.1  | 5  |
| Wild type          | 24.8 $\pm$ 7.4    | 5  |

Serum FSH values were determined as described<sup>44</sup> using radioimmunoassay reagents obtained from the National Hormone and Pituitary Distribution Program, NIDDK. All assays were run in duplicate. <sup>125</sup>I-labelling of FSH was done as described<sup>44</sup>, except that incubation time was reduced to 1 min.

placenta<sup>18,34</sup> and had suggested that it played a major regulatory role in placental hCG production<sup>20</sup>. Our findings that inhibin-deficient mice progress through embryogenesis would normally suggest that inhibin is not essential for extraembryonic development. However, because the mother is a heterozygote, the decidua (the maternal component of the extra-embryonic tissues) may still have the ability to furnish inhibin for placental function.

Every inhibin-deficient male and female mouse eventually develops mixed or incompletely differentiated gonadal stromal tumours<sup>31–33</sup> which may be bilateral and multifocal in origin (Table 1). Because these tumour cells are derived from homologous<sup>31</sup> stromal cell types of the primitive gonadal mesenchyme, it is not surprising that these tumours in both mice and humans<sup>31–33</sup> may take on a mixed appearance with granulosa-like structures in the male and tubular or cord-like structures in the female. In an extensive study of human granulosa cell tumours, 25% of the tumours contained tubular or cord-like structures in addition to the typical granulosa structures<sup>31</sup>.

In the inhibin-deficient mice, tumours are already present both in males and in females at the earliest time point examined (4 weeks), and at that point, the tumours are focal without disruption of gametogenesis or the architecture of the remainder of the gonads (Fig. 4a–d). None of the testes showed Leydig cell hyperplasia in the non-tumour-affected interstitial regions and in fact there is a decrease in Leydig cells with progression of the tumours (data not shown). But over the next few weeks, these tumours become much more destructive, with extensive disruption of normal tissue architecture and aggressive blood vessel invasion even by 5 weeks of age (Table 1, and Fig. 4e–g), leading to the arrest of spermatogenesis and folliculogenesis. Even in the contralateral gonads of some male mice, where tumours have not appeared, there is an arrest of germ-cell maturation. This suggests that the tumours are secreting a substance(s) which may be toxic for spermatogenesis and/or are altering steroid and gonadotropin levels. Because mature-appearing oocytes and spermatozoa were initially present in homozygote  $\alpha$ -inhibin-deficient animals, inhibin does not seem to be necessary for the formation of mature germ cells.

The development of gonadal tumours in these inhibin-deficient mice demonstrates that inhibin plays important autocrine (on Sertoli and granulosa cells) and possibly paracrine (on thecal cells) roles in the gonads as a tumour-suppressor



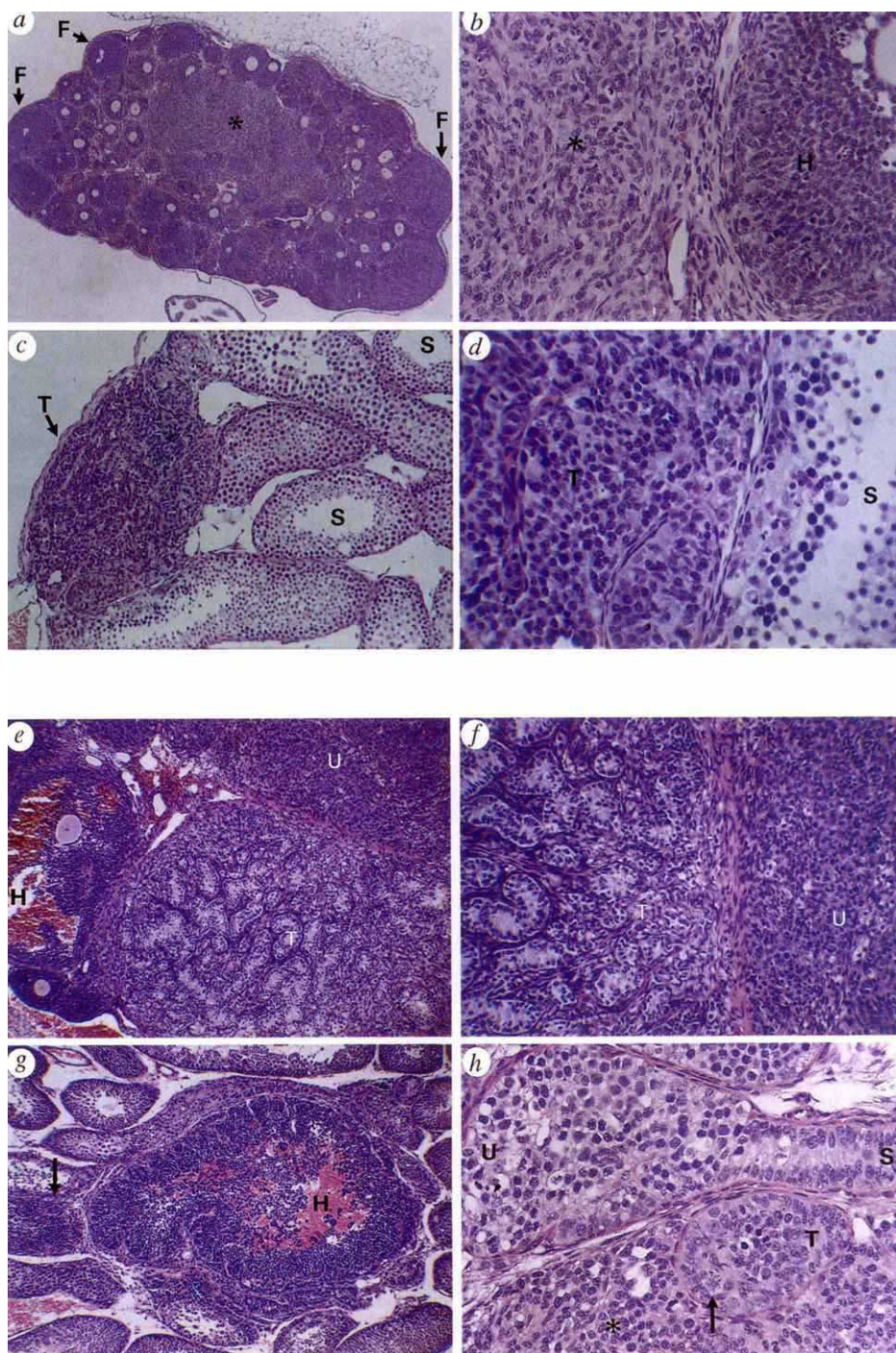
protein. Because primary tumour development in these mice seems to be restricted to the gonads, this result represents the first identification of a tumour-suppressor gene (the  $\alpha$ -inhibin gene) with a gonadal specificity. An analogous situation is seen for many other tumour-suppressor genes (with the exception of p53) which are also very tissue-specific with respect to the affected tissue, at least for the primary neoplasm<sup>22</sup>. Tumour-suppressor proteins have been discovered that appear to act at varying positions in the signal-transduction cascade<sup>22</sup>. The finding here that inhibin-deficient mice develop gonadal tumours demonstrates that inhibin represents the first example of a

heterodimeric tumour-suppressor protein which functions as an extracellular ligand (that is, a secreted protein). These results also suggest that downstream molecules in the inhibin signal transduction cascade should also be tumour-suppressor genes (for example, the inhibin receptor).

Ovarian and testicular physiology is a complex process involving both extragonadal (for example FSH and luteinizing hormone (LH)) and intragonadal regulators (steroids and gonadal peptides such as inhibin and activin). Therefore, the role of the elevated serum FSH in the aetiology of gonadal stromal tumour formation is unclear. We have shown that the FSH levels in the

FIG. 4 Histopathology of tumours from inhibin-deficient mice. *a, b*, Ovary, 4 weeks. In the central region of the gonad, there is a nodule of closely packed, actively dividing, small cells consistent with ovarian stromal precursors (\*). The majority of the surrounding follicles (F) have granulosa cell hyperplasia (H) and normal-appearing oocytes, as is evident with higher magnification. (*a*,  $\times 50$ ; *b*,  $\times 200$ ; haematoxylin-eosin (H and E)). *c, d*, Testes, 4 weeks. A nodule of similar stromal precursor cells fills a portion of seminiferous tubules at the periphery of the cortex (arrow or T). The cells are actively dividing and appear poorly differentiated. Normal seminiferous tubules are labelled as S (*c*,  $\times 150$ ; *d*,  $\times 400$ ; H and E). *e, f*, Ovary, 8 weeks. At low power (*e*), three different tissues are apparent. At the left, a cystic follicle is filled with fresh blood, and granulosa cell hyperplasia surround an oocyte (H). The middle and lower right field is filled by a nodule of tubules lined with Sertoli-like cells (T). At the upper right, there are solid nests of small spindle cells identical to those in the 4-week ovary and typical of an undifferentiated gonadal stromal cell component (U). The latter two patterns are better appreciated with higher magnification (*e*,  $\times 100$ ; *f*,  $\times 200$ ). *g*, Testis, 6 weeks. The centre contains one of three separate nodules found microscopically in a testis that was grossly normal except for a small haemorrhage (H). Extension of the proliferating neoplastic cells into a contiguous seminiferous tubule is seen on the left side of the nodule (arrow). Mature-appearing spermatozoa were present in the epididymis (not shown), indicating normal maturation of the gonad ( $\times 100$ ; periodic acid-Schiff stain). *h*, Testis, 5 weeks. A horizontally oriented seminiferous tubule (deplete of germ cells) is lined by Sertoli cells (S) and distended on the left by the intratubular proliferation of undifferentiated gonadal stromal tumour cells (U). In the large nodule below (\*), foci of cystic degeneration among the tumour cells resemble the macrofollicular pattern of juvenile granulosa cell tumour<sup>32,33</sup>. Other tubules cut in cross-section (T) are also filled by the neoplastic cells. The arrow indicates a mitotic figure in this tubule. ( $\times 200$ , periodic acid-Schiff stain).

**METHODS.** Gonads were processed for histology as described<sup>24</sup>. Frozen sections of the tumours revealed plentiful cytoplasmic lipid in the thecal nodules of the ovarian lesions and some of the cells in the testicular nodules.



No alkaline phosphatase activity was demonstrable in the tumour cells, consistent with the morphological identification of these tumours as gonadal stromal tumours<sup>37-39</sup> as opposed to germ-cell-derived tumours.



inhibin-deficient mice are elevated two- or threefold. But, high FSH is not sufficient for tumour formation because the earliest tumours are focal lesions (Fig. 4a–d). This does not rule out that FSH could be partially responsible for the granulosa cell hyperplasia seen in some females. Furthermore, humans who have FSH-secreting adenomas with serum FSH levels roughly 20-fold higher than normal do not develop gonadal stromal tumours<sup>35</sup>. Although classical experiments, in which testes and ovaries were transplanted under the spleen capsule of gonadectomized rats and mice<sup>36,37</sup>, suggested a role of LH and FSH in tumour development, the tumours that develop in these grafting experiments do not seem to be analogous to those described here because of the much slower appearance, less aggressive nature, and the different types of gonadal stromal tumours that develop<sup>36,37</sup>. It is also unclear how high serum FSH levels in post-menopausal women contribute to the increased incidence of ovarian (epithelial) carcinomas in that age group compared with younger women<sup>38</sup>. The only conclusive way to resolve these questions is to generate mice that are deficient in both inhibin and FSH and analyse gonadal stromal tumour development.

Unlike the other tissue-restricted tumour-suppressor genes, which have relied on familial cancer predisposition syndromes and reverse genetic approaches for the isolation of the relevant gene, the approach described here represents the first example of genetic manipulation of the mouse revealing the tumour-suppressor function of a gene. The targeted deletion of the  $\alpha$ -inhibin gene may be relevant to spontaneous gonadal tumours in man. Because gonadal stromal tumours occur relatively infrequently<sup>31–33</sup>, limited genetic or other studies have been done

to determine if loss of the inhibin subunits or genes correlates with or can serve as a marker for tumour formation. Studies of these types of tumours in a number of species (including humans) reveal that in some, but not all, cases 'immunological' inhibin seems to be present<sup>39–41</sup>. However, it is not known if the inhibin secreted by some of these tumours is biologically active. Further genetic studies of similar types of tumours in man will need to be done to determine the frequency of genetic alteration at the  $\alpha$ -inhibin locus or deletion of the  $\alpha$ -inhibin locus on human chromosome 2q33-qter (mouse chromosome 1)<sup>42</sup>. Although the tumours in the inhibin-deficient mice seem to be restricted to the gonads, this does not rule out that alterations of the  $\alpha$ -inhibin allele may result in progression of other tumours. For example, loss of heterozygosity of chromosome 2q in humans was seen in adenocarcinomas but not squamous cell carcinomas of the lung<sup>43</sup>. In addition, future studies should also examine other inhibin/activin subunit genetic loci (for example  $\beta A$  and  $\beta B$ ) and genetic loci for genes downstream of inhibin in the signal-transduction cascade (such as the inhibin receptor) to understand their role in gonadal tumour development.

The generation of mice deficient in  $\alpha$ -inhibin has identified  $\alpha$ -inhibin as a tumour-suppressor gene and inhibin as a tumour-suppressor protein. The  $\alpha$ -inhibin-deficient mice described here are a useful animal model for understanding gonadal stromal tumour formation as well as the role of inhibin and other factors in formation of these tumours. The inhibin-deficient mice will be especially useful to test the effects of exogenous recombinant inhibin and other antitumour drugs in the suppression of gonadal stromal tumour formation *in vivo*. □

Received 13 August; accepted 23 October 1992.

- Vale, W., Hsueh, A., Rivier, C. & Yu, J. In *Peptide Growth Factors and their Receptors II* (eds Sporn, M. B. & Roberts, A. B.) 211–248 (Springer, Berlin, 1990).
- Smith, J. C., Price, B. M. J., Van Nimmen, K. & Huybroeck, D. *Nature* **345**, 729–731 (1990).
- Thomsen, G. et al. *Cell* **63**, 485–493 (1990).
- Mitrani, E. et al. *Cell* **63**, 495–501 (1990).
- Roberts, A. B. & Sporn, M. B. In *Peptide Growth Factors and their Receptors I* (eds Sporn, M. B. & Roberts, A. B.) 419–472 (Springer, Berlin, 1990).
- Cate, R. L. et al. In *Peptide Growth Factors and their Receptors II* (eds Sporn, M. B. & Roberts, A. B.) 179–210 (Springer, Berlin, 1990).
- Lyons, K. M., Jones, C. M. & Hogan, B. L. M. *Trends Genet.* **7**, 408–412 (1991).
- Massague, J. *Cell* **69**, 1067–1070 (1992).
- Franchimont, P. et al. *Acta endocr.* **98**, 312–320 (1981).
- van Dissel-Emiliani, F. M. F., Grootenhuys, A. J., de Jong, F. H. & de Fooij, D. G. *Endocrinology* **125**, 1899–1903 (1989).
- Mather, J. P. et al. *Endocrinology* **127**, 3206–3214 (1990).
- Bhasin, S. et al. *Endocrinology* **124**, 987–991 (1989).
- Woodruff, T. K. et al. *Endocrinology* **130**, 871–881 (1992).
- Hsueh, A. J. W. et al. *Proc. natn. Acad. Sci. U.S.A.* **84**, 5082–5086 (1987).
- O. W.-S., Robertson, D. M. & deKretser, D. M. *Molec. cell. Endocr.* **62**, 307–311 (1989).
- Woodruff, T. K. et al. *Endocrinology* **127**, 3196–3205 (1990).
- Rabinovici, J., Spencer, S. J. & Jaffe, R. B. *J. clin. Endocr. Metab.* **71**, 1396–1398 (1990).
- Meunier, H., Rivier, C., Evans, R. M. & Vale, W. *Proc. natn. Acad. Sci. U.S.A.* **85**, 247–251 (1988).
- Roberts, V. J., Sawchenko, P. E. & Vale, W. *Endocrinology* **128**, 3122–3129 (1991).
- Petragnola, F., Vaughan, J. & Vale, W. *Proc. natn. Acad. Sci. U.S.A.* **86**, 5114–5117 (1989).
- Chin, T., Parry, R. L. & Donahoe, P. K. *Cancer Res.* **51**, 2101–2106 (1991).
- Weinberg, R. A. *Science* **254**, 1138–1146 (1991).
- Bradley, A., Hasty, P., Davis, A. & Ramirez-Solis, R. *Biotechnology* **10**, 534–539 (1992).
- Donehower, L. A. et al. *Nature* **356**, 215–221 (1992).
- Mansour, S. L., Thomas, K. R. & Capecci, M. R. *Nature* **336**, 348–352 (1988).

- Su, J.-G. J. & Hsueh, A. J. W. *Biochem. biophys. Res. Commun.* **186**, 293–300 (1992).
- McMahon, A. & Bradley, A. *Cell* **62**, 1073–1085 (1990).
- Soriano, P., Montgomery, C., Geske, R. & Bradley, A. *Cell* **64**, 693–702 (1991).
- Kuehn, M. R., Bradley, A., Robertson, E. J. & Evans, M. J. *Nature* **326**, 295–298 (1987).
- Ramirez-Solis, R. et al. *Analyt. Biochem.* **201**, 331–335 (1992).
- Teilmann, G. *Special Tumors of Ovary and Testis* (Lippincott, Philadelphia, 1976).
- Scully, R. E. *Tumors of the Ovary and Maldeveloped Gonads* (Armed Forces Institute of Pathology, Washington DC, 1979).
- Young, R. H. & Telford, A. *Semin. Diag. Path.* **4**, 342–360 (1987).
- Mayo, K. E. et al. *Proc. natn. Acad. Sci. U.S.A.* **83**, 5849–5853 (1986).
- Galway, A. B. et al. *J. clin. Endocr. Metab.* **71**, 907–912 (1990).
- Twombly, G. H., Meisel, D. & Stout, A. P. *Cancer* **2**, 884–892 (1949).
- Gardner, W. V. *Cancer Res.* **15**, 109–117 (1955).
- Hamilton, T. C. *Curr. Probl. Cancer* **16**, 1–57 (1992).
- Lappohn, R. E. et al. *New Engl. J. Med.* **321**, 790–793 (1989).
- De Jong, F. H. et al. *J. Steroid Biochem. molec. Biol.* **37**, 868–866 (1990).
- Piquette, G. N. et al. *Biol. Reprod.* **43**, 1050–1057 (1990).
- Barton, D. E. et al. *Genomics* **5**, 91–99 (1989).
- Tsuchiya, E. et al. *Cancer Res.* **52**, 2478–2481 (1992).
- Keene, J. L., Matzuk, M. M. & Boime, I. *Molec. Endocr.* **3**, 2011–2017 (1989).
- Bradley, A. In *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach* (ed. Robertson, E. J.) 113–151 (IRL, Oxford, 1987).

ACKNOWLEDGEMENTS. We thank J. D. Wallace, S. Jackson and B. Smith for technical assistance, I. Damjanov for review of the histology and suggestions for further analysis of the tumours, W. Moyle and R. Myers for help in the generation of the  $\alpha$ -inhibin cDNA, R. Behringer, L. Donehower and W. Moyle for review of the manuscript and S. Perez and L. Matzuk for manuscript preparation. FSH RIA reagents were a gift from the National Hormone and Pituitary Distribution Program, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and the University of Maryland School of Medicine. This research was supported by grants from the NIH (to M.M., A.B. and A.H.) and the Lalor Foundation (M.M.).



# Fast pulsations in a Wolf-Rayet star

A. Blecha, G. Schaller & A. Maeder

Observatoire de Genève, Chemin des Maillettes 51,  
CH-1290 Sauverny, Switzerland

IN 1930 Eddington suggested<sup>1</sup> that pulsations in stars, particularly in Cepheid variables, might be due to a feedback between core temperature and nuclear energy generation rate, which he called the  $\epsilon$ -mechanism. However, Cepheid pulsation, with a typical period of a few days, is now known to be due to a feedback between temperature and opacity, the  $\kappa$ -mechanism<sup>2</sup>. It has been suggested that Wolf-Rayet stars, the stripped cores of highly evolved stars, might be susceptible to instability by the  $\epsilon$ -mechanism on a characteristic timescale of tens of minutes<sup>3–7</sup>, and that this instability might drive powerful stellar winds which could be responsible for huge mass-loss; Wolf-Rayet stars with initial masses of  $>30 M_{\odot}$  nearly evaporate in their entirety<sup>8</sup>. Here we report the detection of a 627-second periodicity in the WN8 star HD96548 (also known as WR40; ref. 9). Over an observing period of 150 minutes a peak-to-peak variation of 0.005 magnitude was established, and there is some evidence of a first overtone. Theoretical models of WR structure indicate that this periodicity is consistent with the fundamental radial mode, and that it could be partly driven by the  $\epsilon$ -mechanism.

In the  $\epsilon$ -mechanism, the thermodynamic cycle driving the pulsation can be described as follows. Heat is generated during the contraction of the core, thus increasing the temperature and therefore the rate of nuclear energy generation,  $\epsilon$ . The release of extra energy in turn produces expansion above the equi-

lium configuration, contraction then follows and the cycle goes on. This  $\epsilon$ -mechanism does not apply to any known pulsating stars, because the radiative loss (radiative damping) always overcomes the nuclear driving; in fact, it has never been observed. The pulsations of Cepheids and other variable stars are explained by a valve mechanism related to stellar opacities, the  $\kappa$ -mechanism<sup>2</sup>.

Wolf-Rayet stars are considered to be bare cores left over from the peeling of initially massive stars by stellar winds<sup>10</sup>. Analyses of the vibrational stability of WR stars predict instabilities due to the  $\epsilon$ -mechanism in those stars with no hydrogen left. The fundamental radial mode is the most unstable one. The reason for this instability is that the energy released from the large core overcomes the damping in the outer layers so that the overall energy balance during the cycle becomes positive. In WR stars still having some hydrogen at the stellar surface (late WN stars), analyses<sup>11</sup> of models based on the new opacities by Rogers and Iglesias<sup>12</sup> also show instabilities, but in this case partly due to the  $\kappa$ -mechanism in the outer layers. Nonradial instabilities also due to the  $\epsilon$ -mechanism have been predicted<sup>13,14</sup> in WR stars leading to periods of a few hours. The driving of nonradial oscillations by nuclear energy is very unfavourable, because the amplitudes of temperature and radius variations decrease near to the stellar centre and no efficient pumping of energy can occur there<sup>15</sup>.

Detection of the pulsations is hampered by the fact that WR stars generally have extended optically thick winds which do not allow the star itself to be seen. Moffat and Haupt<sup>16</sup> included HD96548 in their sample of WR stars surveyed in narrow band (10–20 nm) filters centred on the line-emission region (H II and N V, III) and continuum at 517 nm. The observations set the upper limit on short-term variation of the continuum to 0.0035 mag (one-third of the observed mean error). More recent

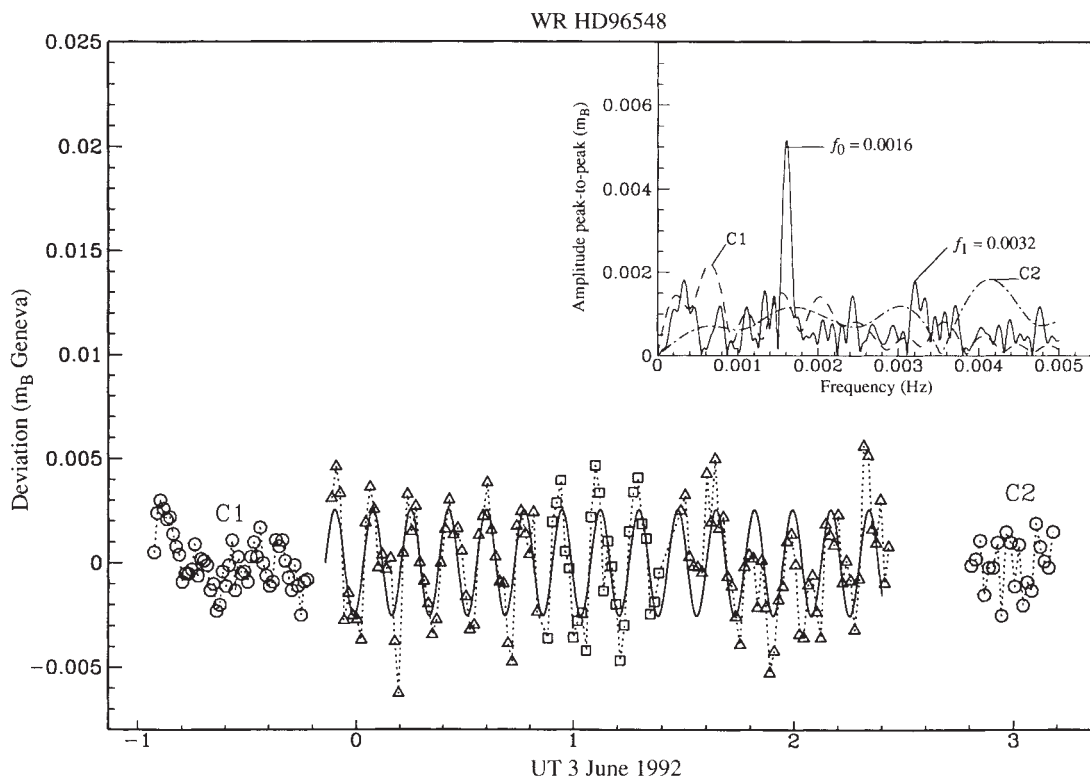


FIG. 1 Observed and reconstructed signal of the WR star. Different symbols are used to distinguish between the three runs of the observation. Each run was made at a different position of the star on the CCD detector. Total number of measurements is 128. The dotted curve joins observed points; the full line represents the reconstructed signal of the fundamental mode.

For comparison the observations of standard stars before (C1) and after (C2) are shown. Inset: the Fourier power spectrum of the observed signal as obtained by the Deeming<sup>25</sup> method. The low-frequency spike close to zero is probably due to the slow variation of the atmospheric transmission. The superimposed dotted line shows the spectra of comparison stars.

observation of a WO (oxygen-rich) star by Lamontagne and Moffat<sup>17</sup> does not reveal a significant oscillation. The largest peak in the Fourier spectrum does not exceed 0.002 mag. Other observations<sup>18,19</sup> of HD96548 have shown variabilities with a possible bi-periodicity of a few days, too long to be explained by the  $\varepsilon$ -mechanism.

In 1991 we started a systematic survey of WR stars using the photoelectric photometer and 70-cm Geneva telescope at the La Silla European Southern Observatory. No oscillation has been observed using the standard observing mode. The telescope has recently been equipped with a CCD camera using the thick, front-illuminated, ultraviolet-coated GEC P8603 416×578 device<sup>20</sup>. Besides the usual imager mode, the CCD camera can be operated as a multichannel photometer (MCP mode). In the MCP mode the CCD is read out according to the observer-specified pattern, thus allowing a fast readout of 1–20 small areas of interest. A series of several hundred observations may be accumulated without data storage and manipulation becoming prohibitive. Although all CCD data are conserved after readout as a mosaic of small images to allow repeated processing, the acquisition system allows for real-time data reduction, permitting the instantaneous display of the photometric parameters of the observed object. The light curve of HD96548 (visible magnitude  $m_V = 7.693$ ) observed on 3 June 1992 is shown in Fig. 1. The magnitude has been corrected for the atmospheric extinction and detector nonuniformity, and the mean magnitude of the observed run was subtracted. Each point is the result of 60-s integration with the Geneva B filter<sup>21</sup> centred at 419 nm. The measured flux is mainly due to the continuum. The contribution of the emission lines He II and N V, III at the filter edge (463 nm) is strongly attenuated by the low transmission of the B filter at that wavelength (30% of the central transmission). The readout area was the rectangle of  $28 \times 28$  arcsec. Unfortunately no comparison star was available within the field of the CCD camera, so only one channel was used. The flux was obtained by integration within the aperture of 20 arcsec centred on the star. The sky background was determined individually for each measurement as a median of the smoothed intensity distribution within the region outside of the aperture. Points with intensities more than  $2.5\sigma$  from the median were removed and the operation repeated until no new points were rejected (sigma clipping). To check for possible instrumental effects, we varied the aperture diameter between 12 and 30 arcsec. This affected the signal noise, but the amplitude and frequency of the oscillation remained unchanged.

The air masses varied between 1.2 and 1.5. Observing was stopped after 128 measurements covering  $\sim 15$  periods, when the scintillation became too strong. A stable standard star was observed for  $\sim 30$  min before and after the run to check the stability of instrumentation and atmospheric conditions.

The Fourier analysis of the light curve reveals, not surprisingly, a single frequency of  $f_0 = 0.0016 \pm 0.00006$  Hz corresponding to the period of 627 s. When examining the power spectrum (inset to Fig. 1), we see that its amplitude peak-to-peak, 0.005

mag, is well above the background noise. The r.m.s. deviation of the light curve of the WR star from the mean was 0.0025 mag. In comparison the r.m.s. deviation from the mean of standard star was 0.0012 mag both before and after the observation of the WR star. The r.m.s. of the residual noise is 0.0012 magnitude.

We now compare the observations to the model predictions<sup>6,8</sup> which indicate pulsational instabilities in WR stars with masses above  $12\text{--}16 M_\odot$ . Periods predicted for the radial fundamental mode ranges between 15 and 60 min, for models with initial masses above  $40 M_\odot$ . The expansion of the star leads to the variation of the surface temperature and of the luminosity. We must however specify the period corresponding to the case of WR40 which has a mass slightly lower than that considered in the models. Nonradial pulsations with typical periods of several hours can be excluded. There is no evidence for any frequency lower than  $f_0$  except for the peak close to the zero frequency, probably caused by the variation of the atmospheric extinction. The barely visible spike in the power spectrum at  $f_1 = 0.0032$  Hz may be interpreted as an overtone of the  $f_0$ , although it could be due merely to a nonsinusoidal form of the light curve (non-linear pulsations). If the peak at  $f_1 = 0.0032$  Hz is due to an overtone, the frequency  $f_0$  corresponds to the fundamental mode, as Schaller<sup>11</sup> found that the theoretical frequency ratios for  $f_1/f_0$  were 0.45–0.55 and those for  $f_2/f_1$  were 0.74–0.76. An estimate using the existing models shows that the observed period is in agreement with a general parameter describing the whole category of WR stars: the pulsation constant  $Q$  defined by the relation  $Q = P[M/M_\odot/(R/R_\odot)^3]^{1/2}$ , where  $R_\odot$  is the solar radius. The value of  $Q$  is 0.04 and essentially constant during the WR phase<sup>4</sup>. An intrinsic luminosity of  $\log L/L_\odot = 5.4$  (where  $L_\odot$  is solar luminosity) has been obtained<sup>22</sup> from spectral analysis of HD96548, and this is also constant during the WN stage at a given mass. The mass  $M$  and the radius  $R$  can be estimated, using the relations between  $M$ ,  $R$  and the luminosity of WR stars obtained<sup>23</sup> from a recent grid of stellar models<sup>24</sup>. We estimate that the actual mass of WR40 is  $14\text{--}15 M_\odot$ . In this case the initial mass would have been about  $29.5 M_\odot$  and the actual radius of the hydrostatic core between  $0.75$  and  $0.95 R_\odot$ . The pulsation constant  $Q$  corresponding to the observed period is then between 0.030 and 0.043, in good agreement with the value found for the fundamental radial mode in existing models.

The reality of pulsations in WR40 seems well established. Comparison of observed and theoretical pulsations also supports the identification of pulsations in the fundamental mode. But until we have better data on the chemical composition of WR40, we cannot say with certainty whether the phenomenon responsible for the pulsation is the  $\varepsilon$ - or  $\kappa$ -mechanism, or a combination of the two. We do not know how often the cloud surrounding the WR star is sufficiently transparent to allow the pulsations to be observed. Further monitoring of WR40 and other WR stars will help us to understand these unstable stellar cores, and the possible relation between pulsations and the huge mass loss rates observed for these objects.  $\square$

Received 27 July; accepted 16 October 1992.

1. Eddington, A. S. *The Internal Constitution of the Star* (Cambridge Univ. Press, 1930).
2. Cox, J. P. & Giuli, R. T. *Principles of Stellar Structure*, Vol. 2, 1106–1122 (Gordon and Breach, New York, 1968).
3. Noels, A. & Gabriel, M. *Astr. Astrophys.* **101**, 215–222 (1981).
4. Maeder, A. *Astr. Astrophys.* **147**, 300–308 (1985).
5. Cox, A. N. & Cahn, J. H. *Astrophys. J.* **326**, 804–812 (1988).
6. Maeder, A. & Schaller, G. in *IAU Symp. No. 143 WR Stars and Interrelations with Other Massive Stars in Galaxies* (eds van der Hucht, K. & Hidayat, B.) 167–175 (Kluwer, Amsterdam, 1991).
7. Schaller, G. *ASP Conf. Series, Confrontation Between Stellar Pulsation and Evolution* **11** (eds Cacciati, C. & Clementini, G.) 304–307 (Bookcrafters, San Francisco, 1991).
8. Maeder, A. *Q. J. R. astr. Soc.* **32**, 217–231 (1991).
9. van der Hucht, K. A., Conti, P. S., Lundström, I. & Stenholm, B. *Space Sci. Rev.* **28**, 227–306 (1981).
10. Lamers, H., Maeder, A., Schmutz, W. & Cassinelli, J. P. *Astrophys. J.* **368**, 538–544 (1991).
11. Schaller, G. thesis, Univ. Geneva (1992).
12. Rogers, F. J. & Iglesias, C. A. *Astrophys. J. Ser. S* **79**, 507–568 (1992).
13. Kiribay, H., Bertelli, G. & Chiosi, C. in *Theoretical Problems in Stellar Stability and Oscillations*, *Proc. 25th Liège int. AP. Colloquium*, 126–134 (Liège, Université de Liège, 1984).

14. Noels, A. & Scuflaire, R. *Astr. Astrophys.* **161**, 125–129 (1986).
15. Simon, R. *Bull. Acad. R. Belgique, Cl. Sci.* **43**, 610–618 (1957).
16. Moffat, A. F. J. & Haupt, W. *Astr. Astrophys.* **32**, 435–439 (1974).
17. Lamontagne, R. & Moffat, A. F. J. *Astr. Astrophys.* **162**, 114–116 (1986).
18. Gosset, E. & Vreux, J.-M. *Astr. Astrophys.* **231**, 100–134 (1990).
19. Balona, L. A., Egan, J. & Marang, F. *Mon. Not. R. astr. Soc.* **240**, 103–115 (1989).
20. Blecha, A., Weber, L., Simond, G. & Quille, D. in *CCDs in Astronomy, A.S.P. Conf. Series* **8**, 192–200 (1990).
21. Rufener, F. & Nicolet, B. *Astr. Astrophys.* **206**, 357–374 (1988).
22. Hamann, W.-R., Dünnebell, G., Koesterke, L. & Wessolowski, U. *Astr. Astrophys.* **249**, 443–454 (1991).
23. Schaerer, D. & Maeder, A. *Astr. Astrophys.* **263**, 129–136 (1992).
24. Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. *Astr. Astrophys. suppl.* (in the press).
25. Deeming, T. J. *Astrophys. Space Sci.* **36**, 137–158 (1974).

ACKNOWLEDGEMENTS. We thank the observatory staff involved in the development of the CCD acquisition system, especially L. Weber and G. Simond, and P. Bartholdi for the SuperMongo procedures for Fourier analysis. This work was partly supported by the Swiss Fonds National de la Recherche Scientifique.



# A 600-day periodicity in solar coronal holes

P. S. McIntosh, R. J. Thompson\* & E. C. Willock\*

Space Environment Laboratory, 325 Broadway, Boulder, Colorado 80303, USA

\* IPS Radio and Space Services, PO Box 5606, West Chatswood, New South Wales 2057, Australia

**CORONAL** holes are large-scale structures of lower density and temperature in the solar corona. They appear as large dark features in X-ray and radio images, and as bright areas in infrared He I images of the Sun. The longest series of data is the He I (10,830 Å) observations published within H $\alpha$  synoptic charts<sup>1</sup>. Here we analyse outlines of coronal holes from these charts to show the variation of coronal hole area over the period 1977–89. We find that the total area of coronal holes enclosed within a latitude band of 10–50° S shows a 592-day periodic variation throughout the 13-year interval, which includes all of solar cycle 21 and the first three years of cycle 22. The corresponding northern latitude region does not show this periodic behaviour. There is some indication of an association between the coronal hole variation and other solar phenomena such as the occurrence of super-active regions and the reported 153-day periodicity.

Synoptic charts<sup>1–4</sup> from H $\alpha$  images map the heliographic positions of filaments, filament channels, large sunspots, chromospheric plage and magnetic polarity in charts covering the entire solar globe for each 27-day solar rotation. Coronal hole outlines, inferred from observations in He I (10,830 Å), were incorporated into the charts from December 1976. This places coronal holes in context with patterns of large-scale solar magnetic fields, filaments and active regions, and provides a ready means to follow the long-term evolution of these important features. The database analysed here is internally consistent, and is the longest time series of coronal hole data yet analysed.

Coronal hole data within the synoptic maps is based on He I (10,830 Å) images in which coronal holes appear as bright areas lacking the dark network structures of the normal, quiet chromosphere and transition zone<sup>5</sup>. Evolution of these holes appears systematic<sup>6</sup>, and consistent with the limited data on X-ray coronal holes obtained from rockets and orbiting spacecraft<sup>7–9</sup>. When deriving the coronal hole outlines for inclusion in the synoptic charts, one of us (P.S.M.) examined high-quality photographic He I (10,830 Å) images for each day to determine the most consistent position and shape of each coronal hole, using solar rotation and the consistency among the images to subjectively determine the presence of the coronal hole.

Digital versions of H $\alpha$  synoptic charts for 1977–87 were analysed at IPS Radio and Space Services to determine the behaviour of coronal hole area. To extend the data interval, coronal hole areas for 1988–89 were measured directly from the published analogue synoptic charts<sup>1</sup>. Figure 1 shows the variation of the area of solar coronal holes for northern and southern latitude bands from solar Carrington rotation (CR) 1649 (December 1976) to CR1822 (December 1989), an interval including nearly all of solar cycle 21 and more than 3 years of cycle 22. The area of the southern coronal holes shows a periodic variation with eight nearly equally spaced peaks at intervals of ~600 days. The amplitude of the peaks changes considerably; three large peaks occur between September 1980 and April 1985, corresponding to the late-maximum and the declining phase of the sunspot cycle. The last two peaks in the sequence (at CR1792 and CR1811), both during cycle 22, are less distinct and are intermixed with smaller peaks at CR1802 and, possibly, at CR1820. The northern area peaks, in contrast, show no simple periodic variation but do show a similar general increase in amplitude in the declining phase of cycle 21.

Figure 2 is a 'stackplot' for southern coronal holes. It is formed

by taking the latitude slice of 10° S to 50° S from each digital synoptic map and arranging them in a vertical stack to display the evolution of coronal holes. To improve the visibility of the long-lived structures, the vertical scale has been compressed and three identical stackplots have been placed side by side with a one-rotation offset to show the drift of features across the boundaries of adjacent plots. The coronal holes (shaded black) show areal variations corresponding to the peaks in Fig. 1b, being especially visible for the large peaks of 1981–84. Figure 2 also demonstrates that the variation in coronal hole area was not the product of an increase in the number of coronal holes; rather the areal peaks were the result of the growth of a single, very large, coronal hole at the times of each peak.

Fourier power spectra (Fig. 3) were determined for the northern and for the southern coronal hole area signals using the autocorrelation method with a Hanning 'raised cosine' as the windowing function. The central peak was suppressed by subtracting the mean value from the signal. The southern area power spectrum has two distinct peaks: a long-period peak at 3,490 days (nearly a solar cycle), related to the envelope of the area variation, and a shorter-period peak near 600 days. The power spectrum for the northern coronal hole area variations shows no peak near 600 days.

Least-squares best-fit equations (1) and (2) give the time of the area peaks for the southern coronal holes in terms of Carrington rotations and years, respectively. The coronal hole peak is taken as the time at which the area averaged over four rotations reaches a maximum, and each peak is numbered sequentially from  $m = 1$  at the first peak (CR1662). The equations predict

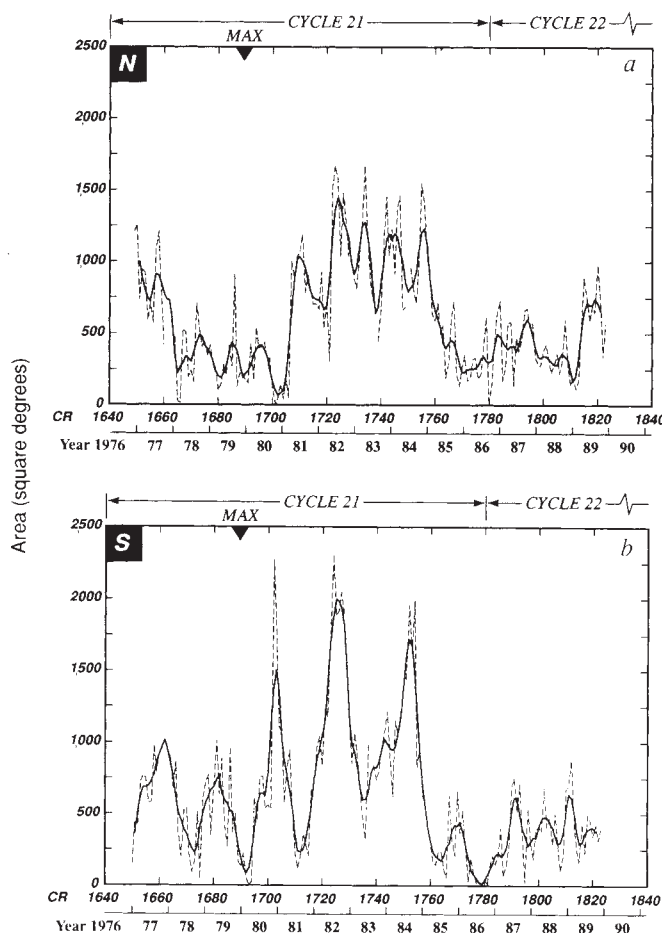


FIG. 1 Variation of coronal hole area in heliographic degrees between CR1649 and CR1822 (December 1976 to December 1989) for (a) northern hemisphere (upper panel) and (b) southern hemisphere (lower panel). Area was integrated in the latitude zone 10–50°. The heavy line is the four-rotation running average, and the thin line indicates area for a single rotation.

the timing of the peaks with a root-mean-square error of 1.9 Carrington rotations or 0.14 years (51 days). From the fit, the period and its standard error are  $21.7 \pm 0.3$  Carrington rotations ( $592 \pm 9$  days, or  $1.622 \pm 0.025$  years).

$$p_{CR}(m) = 1640.0(\pm 1.7) + 21.7(\pm 0.3) m \quad (1)$$

and

$$p_{years}(m) = 1976.26(\pm 0.13) + 1.622(\pm 0.025) m \quad (2)$$

There are three remarkable features of the coronal hole variation: its periodicity, the north-south asymmetry and its long lifetime. All three properties have been reported for other, possibly related, solar phenomena. Periodicities of between 152 and 155 days in the occurrence of solar flares have been suggested<sup>10-13</sup>. This period is almost exactly one-quarter of the period reported here for coronal holes. An association between the two is supported by Thompson and Willock<sup>14</sup>, who suggested that the 153-day periodicity found in solar flares during cycle 21 arose mostly from two flare-productive regions spaced at 612 days (SESC region 2779 in November 1980 and region 3804 in July 1982).

The second remarkable feature of the coronal hole variations—the hemispheric asymmetry—has been reported in a

wide range of solar features (see ref. 15 for a review). Hoeksema and Scherrer<sup>16</sup> have also reported a hemispheric asymmetry in the rotation of the coronal magnetic field modelled using the observed photospheric field during 1976 to 1985. Observations of many other north-south asymmetries indicate a shift to dominance by the southern hemisphere<sup>17-19</sup> since cycle 20, roughly the interval discussed here.

Third, long-lived features have been reported by Svalgaard and Wilcox<sup>20</sup> in the inferred interplanetary magnetic sector structure deduced from geomagnetic disturbances over the interval of 1926-73. Coronal holes, solar wind streams and geomagnetic disturbances are known to be strongly interrelated<sup>21,22</sup>, suggesting that the long-lived features found by Svalgaard and Wilcox may have been associated with coronal holes.

Some of the most notable solar active regions have developed near large coronal holes as they expanded rapidly in area<sup>23,24</sup>. McIntosh<sup>25</sup> has listed the most active regions in the period of 1972-89. Of the 13 regions listed, 11 occurred during the period of the coronal hole observations discussed here. Within a periodicity of almost 22 Carrington rotations, 9 of these 11 regions fell in a band of seven rotations near the coronal hole peaks. If all 11 events are considered as being independent, this situation has a binomial chance probability of  $P = 0.001$ . It is likely, however, that not all the events are independent and therefore that the chance probability is larger. For this reason, the association between the coronal hole peaks and large flaring regions must be regarded as suggestive.

Because of the short interval of the present data, it is not possible to be sure of the duration and persistence of the 600-day coronal hole periodicity. It may be that such periodicities come and go, or shift in phase or hemispheric dominance. Indeed, since the end of the present data set, there is evidence from the occurrence of active solar regions and from geomagnetic disturbances to suggest that such a shift may be under way.

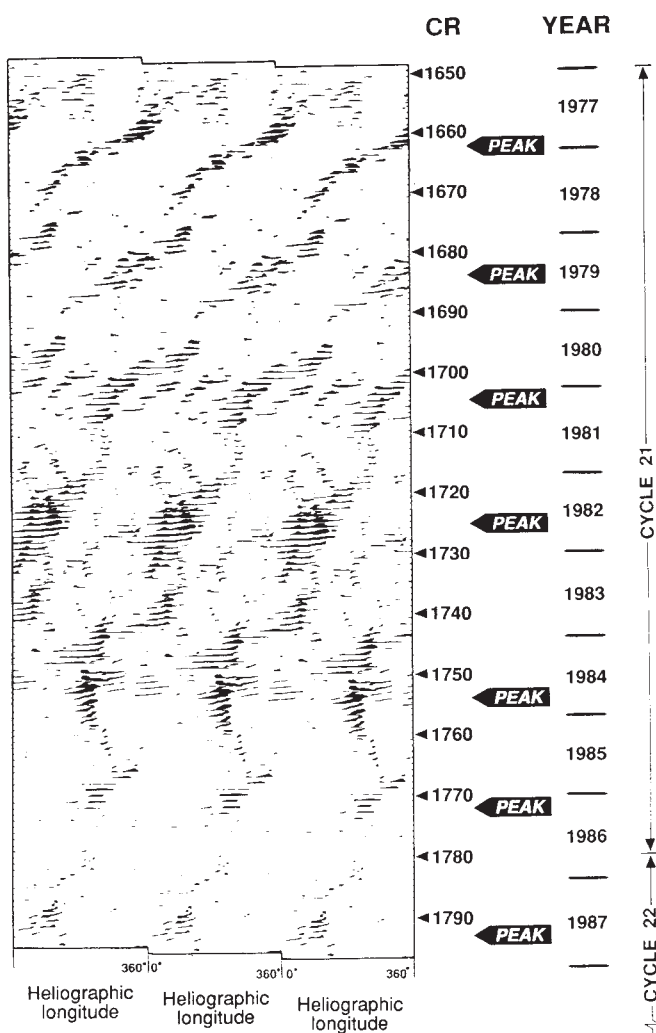


FIG. 2 Southern hemisphere coronal holes for more than a solar cycle (1977-87), displayed in a vertical time series of latitude zones (stackplot) for latitudes  $10^{\circ}$ - $50^{\circ}$  S. Three solar rotations of longitude are placed side by side with a one-rotation offset to permit continuous viewing of long-lived series of coronal holes that drift with respect to heliographic longitude. The coronal holes are shaded dark. Times of maximum coronal hole area from Fig. 1b are indicated at the right.

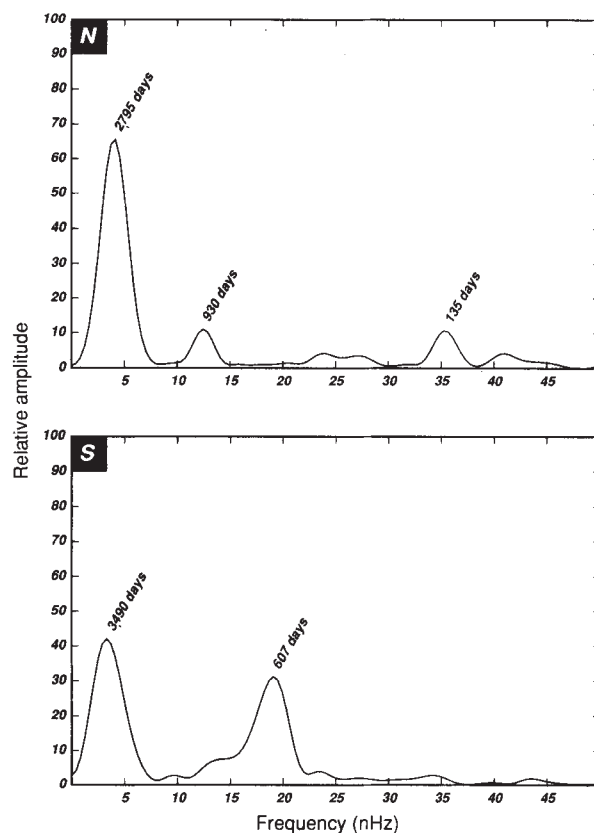


FIG. 3 Fourier power spectra for the area variation of southern hemisphere coronal holes (lower panel) and northern coronal holes (upper panel) between CR1649 and CR1822 (December 1976 to December 1989).



Nevertheless, any solar phenomenon that is both long-lived and periodic may be predictable. The evidence presented here suggests that the occurrence, and perhaps the solar location, of large coronal holes could be predicted many months in advance. The established link between coronal holes and geomagnetic disturbances, and the possible link to super-active regions, raise the possibility of predicting the most important solar-induced disturbances. The existence of stability, periodicity and hemispheric asymmetry in coronal holes implies long-lived, non-random generation of magnetic flux. This will affect our understanding of the most fundamental problems in solar astrophysics, the nature of the solar cycle.  $\square$

Received 20 March; accepted 6 October 1992.

1. *Solar-Geophysical Data* (World Data Center-A for Solar-Terrestrial Physics, Boulder Colorado).
2. McIntosh, P. S. *UAG Rep. 70* (World Data Center A for Solar-Terrestrial Physics, Boulder Colorado, 1979).
3. McIntosh, P. S., in *Solar Activity Observations and Predictions* (eds McIntosh, P. S. & Dryer, M.) 65–92 (Massachusetts Inst. Technol., 1972).
4. McIntosh, P. S. *Rev. Geophys. Space Phys.* **10**, 837–846 (1972).
5. Harvey, J. W., Krieger, A. S., Davis, J. M., Timothy, A. F. & Vaiana, G. S. *Bull. Am. astr. Soc.* **7**, 358 (1975).
6. McIntosh, P. S., Willock, E. C. & Thompson, R. J. *UAG Rep. 101* (World Data Center A for Solar-Terrestrial Physics, Boulder Colorado, 1991).
7. Harvey, J. W. *Osserv. Mem. "Osserv. Astrofis. Arcetri"*, **104**, 50–58 (1975).
8. Harvey, J. W. & Sheeley, N. R. Jr. *Space Sci. Rev.* **23**, 139–158 (1979).
9. Kahler, S. W., Davis, J. M. & Harvey, J. W. *Solar Phys.* **87**, 47–56 (1983).
10. Rieger, E. *et al. Nature* **312**, 623–625 (1984).
11. Ichimoto, K., Kubota, J., Suzuki, M., Tohmura, I. & Kurokawa, H. *Nature* **316**, 422–424 (1985).
12. Bai, T. & Sturrock, P. A. *Nature* **327**, 601–604 (1987).
13. Bogart, R. S. & Bai, T. *Astrophys. J.* **299**, L51–55 (1985).
14. Thompson, R. J. & Willock, E. C. *Proc. Leura Solar Terrestrial Predictions Workshop*, Vol. 1 (eds Thompson, R. J. *et al.*), 603–610 (1990).
15. Garcia, H. A. *Solar Phys.* **127**, 185–197 (1990).
16. Hoeksema, J. T. & Scherrer, P. H. *Astrophys. J.* **318**, 428–436 (1987).
17. Hansen, R. & Hansen, S. *Solar Phys.* **44**, 225–230 (1975).
18. Verma, V. K. *Solar Phys.* **114**, 185–188 (1987).
19. Roy, J.-R. *Solar Phys.* **52**, 53–61 (1977).
20. Svalgaard, L. & Wilcox, J. M. *Solar Phys.* **41**, 461–475 (1975).
21. Sheeley, N. R. Jr., Harvey, J. W. & Feldman, W. C. *Solar Phys.* **49**, 271–278 (1976).
22. Sheeley, N. R. Jr., Harvey, J. W. *Solar Phys.* **70**, 237–249 (1981).
23. McIntosh, P. S. *EOS* **72**, 383 (1991).
24. McIntosh, P. S., in *Proc. Solar Cycle Workshop* (ed. Harvey, K.) (National Solar Observatory, Sunspot, New Mexico, in the press).
25. McIntosh, P. S. *Solar Phys.* **125**, 251–267 (1990).

ACKNOWLEDGEMENTS. We thank S. Suess for converting the analog synoptic maps into digital format, B. Paterson for assistance with Fourier analysis and D. Tranquille for assistance with the figures.

## Phase behaviour of metastable water

Peter H. Poole, Francesco Sciortino, Ulrich Essmann & H. Eugene Stanley

Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215 USA

THE metastable extension of the phase diagram of liquid water exhibits rich features that manifest themselves in the equilibrium properties of water. For example, the density maximum at 4 °C and the minimum in the isothermal compressibility at 46 °C are thought to reflect the presence of singularities in the behaviour of thermodynamic quantities occurring in the supercooled region<sup>1,2</sup>. The 'stability-limit conjecture'<sup>3–5</sup> suggests that these thermodynamic anomalies arise from a single limit of mechanical stability (spinodal line), originating at the liquid–gas critical point, which determines the limit of both superheating at high temperatures and supercooling at low temperatures. Here we present a comprehensive series of molecular dynamics simulations which suggest that, instead, the supercooling anomalies are caused by a newly identified critical point, above which the two metastable amorphous phases of ice (previously shown to be separated by a line of first-order transitions<sup>6,7</sup>) become indistinguishable. The two amorphous ice phases are thus incorporated into our understanding of the liquid state, providing a more complete picture of the metastable and stable behaviour of water.

The equation of state of a fluid expresses the shape of the surface  $P(V, T)$  in the space of pressure  $P$ , specific volume  $V$  and temperature  $T$ . Spinodal lines on the  $P(V, T)$  surface separate metastable states from those that are mechanically unstable. In practice, nucleation of a stable phase precludes study of metastable states arbitrarily close to the spinodal<sup>8</sup>. Many thermodynamic functions and dynamic properties of the system, however, become singular near a spinodal line in a fashion analogous to their behaviour at a critical point<sup>9</sup>. Hence, although the spinodal may not be attainable, its influence on nearby states can still be observed.

One well-known equilibrium anomaly of a liquid is the existence at constant  $P$  of a temperature of maximum density (TMD). As  $P$  varies, the TMD becomes a line, the TMD line.

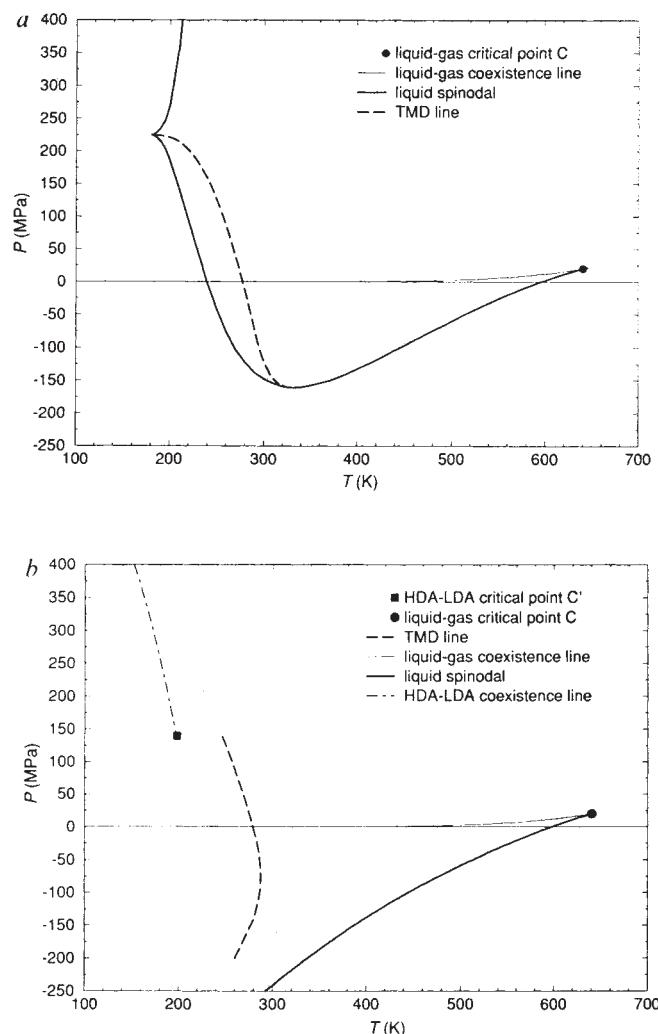


FIG. 1 *a*, The phase diagram predicted by the stability-limit conjecture<sup>4,12</sup>, showing the TMD line intersecting a re-entrant liquid spinodal line. The liquid–gas coexistence curve ends in the liquid–gas critical point C. *b*, Phase diagram of ST2 water. The liquid spinodal originating at the liquid–gas critical point C does not intersect the TMD line. The existence of a gap between the TMD line and the spinodal implies that the anomalies are suppressed by sufficiently negative  $P$ , so that the region between the TMD line and the spinodal is characterized by a normal (positive) isobaric expansivity. The appearance of a TMD line with positive slope has been observed in simulations of the gaussian core model<sup>27</sup>, a simple pair-potential unrelated to ST2. Preliminary results on the TIP4P potential indicate a similar behaviour of the TMD line. A new coexistence line, distinct from the liquid–gas coexistence line, separates two phases analogous in structure to the LDA and HDA ice phases, and ends in a critical point C'. The locations of the liquid spinodal and of C' estimated here take into account the pressure and temperature shifts of the ST2 data described in Fig. 5.

As first argued for water<sup>4</sup>, and later<sup>10–12</sup> for any liquid exhibiting a TMD, there exist thermodynamic constraints on the way a spinodal and a TMD line may meet. The spinodal line  $P_s(T)$  originates at the liquid–gas critical point and lies below the liquid–gas coexistence line in the  $P$ – $T$  phase diagram (Fig. 1a).

Experiments show<sup>13</sup> that  $P_s(T)$  passes into the negative  $P$  region of the phase diagram. Negative  $P$  thermodynamic states are well-defined (metastable) states of hydrostatic tension, so the spinodal line at negative  $P$  represents the limit of tensile strength of the liquid. The observed TMD line of liquid water, where it

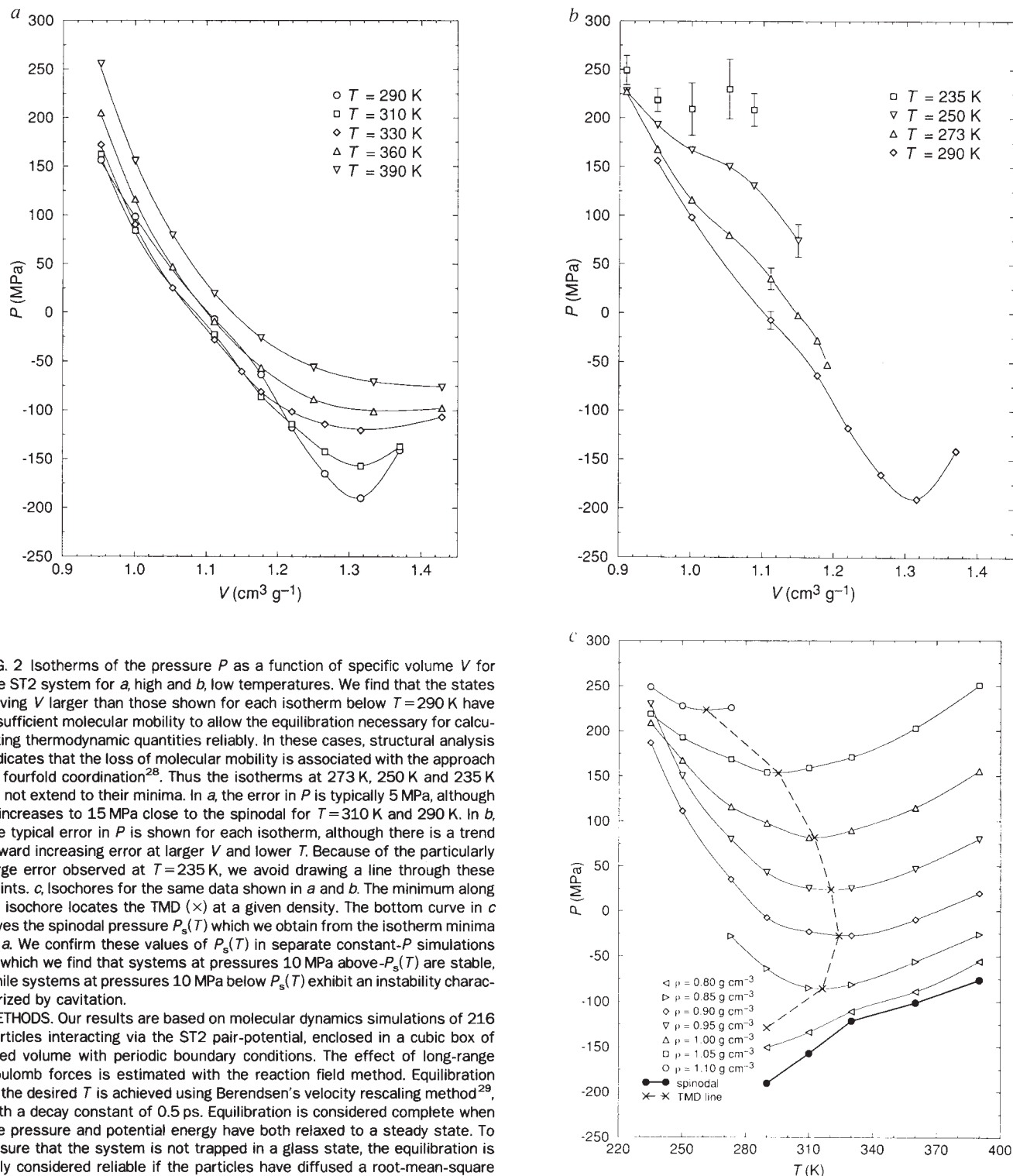


FIG. 2 Isotherms of the pressure  $P$  as a function of specific volume  $V$  for the ST2 system for *a*, high and *b*, low temperatures. We find that the states having  $V$  larger than those shown for each isotherm below  $T = 290 \text{ K}$  have insufficient molecular mobility to allow the equilibration necessary for calculating thermodynamic quantities reliably. In these cases, structural analysis indicates that the loss of molecular mobility is associated with the approach to fourfold coordination<sup>28</sup>. Thus the isotherms at  $273 \text{ K}$ ,  $250 \text{ K}$  and  $235 \text{ K}$  do not extend to their minima. In *a*, the error in  $P$  is typically  $5 \text{ MPa}$ , although it increases to  $15 \text{ MPa}$  close to the spinodal for  $T = 310 \text{ K}$  and  $290 \text{ K}$ . In *b*, the typical error in  $P$  is shown for each isotherm, although there is a trend toward increasing error at larger  $V$  and lower  $T$ . Because of the particularly large error observed at  $T = 235 \text{ K}$ , we avoid drawing a line through these points. *c*, Isochores for the same data shown in *a* and *b*. The minimum along an isochore locates the TMD (x) at a given density. The bottom curve in *c* gives the spinodal pressure  $P_s(T)$  which we obtain from the isotherm minima in *a*. We confirm these values of  $P_s(T)$  in separate constant- $P$  simulations in which we find that systems at pressures  $10 \text{ MPa}$  above  $-P_s(T)$  are stable, while systems at pressures  $10 \text{ MPa}$  below  $-P_s(T)$  exhibit an instability characterized by cavitation.

**METHODS.** Our results are based on molecular dynamics simulations of 216 particles interacting via the ST2 pair-potential, enclosed in a cubic box of fixed volume with periodic boundary conditions. The effect of long-range Coulomb forces is estimated with the reaction field method. Equilibration to the desired  $T$  is achieved using Berendsen's velocity rescaling method<sup>29</sup>, with a decay constant of  $0.5 \text{ ps}$ . Equilibration is considered complete when the pressure and potential energy have both relaxed to a steady state. To ensure that the system is not trapped in a glass state, the equilibration is only considered reliable if the particles have diffused a root-mean-square distance of at least one molecular diameter from their starting positions. The average value of  $P$  is measured in a continuation of the simulation after equilibration, in which  $T$  is controlled with Berendsen's method, and confirmed in further constant- $(N, V, E)$  simulations. The time step for the update of particle positions is fixed at  $1.0 \text{ fs}$ . In almost all cases, a given simulation is started using the final particle positions of a previous simulation at a state point as near as possible to the new one. The lengths of the runs (equilibration plus production) vary depending on the time needed to

obtain equilibration and measurements, but are not less than  $200 \text{ ps}$  except at the highest  $T$ . Because of the slow relaxation found in the systems at the lowest  $T$  and highest specific volumes, run times of  $600$  to  $800 \text{ ps}$  are typical in this regime. To confirm that the behaviour we find near the spinodal is not an artefact of the small system size, we have reproduced the position and shape of the  $\rho = 0.80 \text{ g cm}^{-3}$  isochore shown in *c* in constant  $P$  simulations of a system of  $1,728$  ST2 particles.



has been measured<sup>14</sup>, has negative slope, and would intersect the spinodal line at negative  $P$  if this trend were to continue. But the thermodynamic constraints on the intersection of spinodal and TMD lines demand<sup>10–12</sup> that  $P_s(T)$  passes through a minimum at this intersection point;  $P_s(T)$  moves back to higher  $P$  as  $T$  is further lowered, and eventually becomes positive at sufficiently low  $T$  (Fig. 1a). The stability-limit conjecture states that such a 're-entrant' spinodal actually occurs in liquid water, and moreover gives rise to the anomalous behaviour of this liquid at low  $T$ . Note that this hypothesis requires that a spinodal line  $P_s(T)$  always has negative slope in the  $T$  range of the phase diagram in which a TMD line is found.

We initially sought to discover the extent to which the stability-limit conjecture (represented by Fig. 1a) is borne out in molecular dynamics computer simulation, but in fact we find that our results suggest a rather different phase diagram (Fig. 1b). In the present simulations, we use the ST2 potential<sup>15</sup> to calculate the interparticle interactions because, out of the potentials that we tested (TIP4P<sup>16</sup>, SPC/E<sup>17</sup> and ST2), ST2 exhibits the best developed density maximum. Additionally, the range of temperature between the TMD and the lowest temperature that can be studied (beyond which there is insufficient molecular mobility) is much larger for ST2 than for the other potentials. Thus ST2 water can be studied deep into the anomalous region.

We locate the spinodal  $P_s(T_0)$ , for a fixed temperature  $T_0$ , from the  $T = T_0$  isotherm in a plot of  $P$  against  $V$  as the point for which the pressure is a minimum (Fig. 2a). This is equivalent to the condition that the isothermal compressibility  $K_T = -1/V(\partial V/\partial P)_T$  becomes singular.

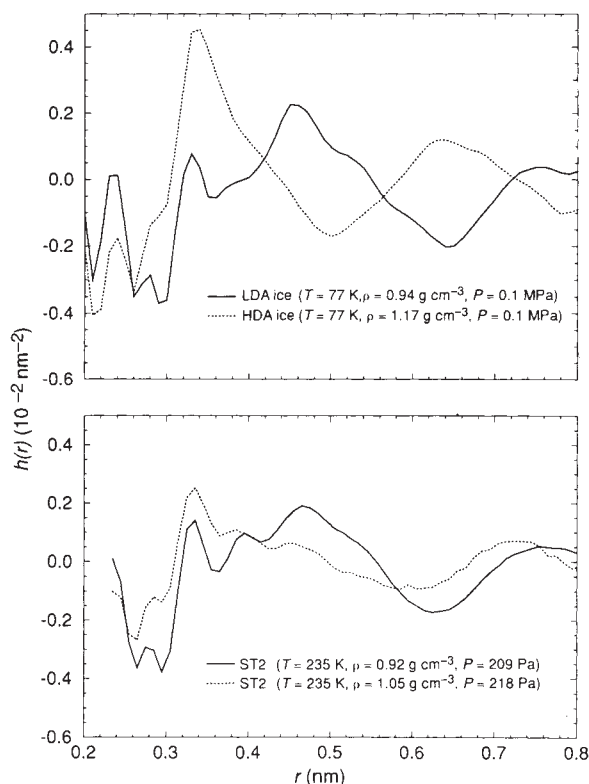


FIG. 3 Plots of the neutron-weighted pair correlation function  $h(r)$  as determined in experiment<sup>19</sup> at  $T = 77 \text{ K}$  for the HDA and LDA ices (top panel) compared with the ST2 system at  $235 \text{ K}$  at two values of density straddling the critical density (bottom panel). Here  $h(r)$  is defined in terms of the radial distribution functions  $g_{\text{O-O}}(r)$ ,  $g_{\text{O-H}}(r)$  and  $g_{\text{H-H}}(r)$ , which give the probabilities that an atom of a particular type is in a differential radial distance element  $dr$  at a distance  $r$  from a given atom:  $h(r) = 4\pi\rho(0.092g_{\text{O-O}}(r) + 0.422g_{\text{O-H}}(r) + 0.486g_{\text{H-H}}(r) - 1)$ , where the coefficients are appropriate for neutron scattering experiments<sup>30</sup>, O refers to oxygen, and H to hydrogen.

Unexpectedly, we find that the spinodal is a monotonic function of  $T$  and thus is not re-entrant (a result confirmed independently<sup>18</sup>), even at temperatures below those at which a TMD is observed (Fig. 2c). Such behaviour, although different from the stability-limit conjecture, is nonetheless thermodynamically consistent. The TMD line, although it has negative slope at high  $P$ , acquires a positive slope at low  $P$ , thereby avoiding an intersection with the spinodal, and thus pre-empting a re-entrant spinodal. The assumption made in the stability-limit conjecture, that the TMD line has negative slope throughout the phase diagram, is therefore not borne out by our calculations (Fig. 2c).

If there is no re-entrant spinodal, then what causes the known anomalous behaviour of water? Our calculations suggest that the cause is a novel critical point  $C'$ , which is unrelated to the liquid-gas critical point  $C$ . The evidence is four-fold:

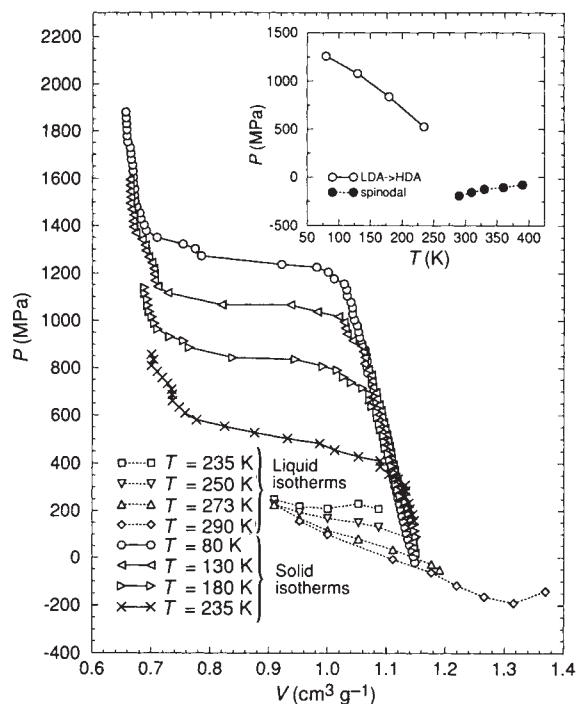


FIG. 4 Isothermal compression curves for the low-density amorphous solid ST2 system. For comparison, the low-temperature ST2 liquid isotherms are reproduced from Fig. 2b. We simulate the isothermal compression employed in the experiments by monitoring the specific volume of the system during a simulation in which  $T$  and  $P$  are controlled with Berendsen's method, and in which  $P$  is increased stepwise. A similar simulation procedure<sup>31</sup> has been successful in reproducing the experimentally observed  $P$ -induced transition from ice  $I_h$  to HDA ice. We apply this procedure to LDA ice obtained by cooling the simulated low density liquid to below  $T_c$ . The inset shows the pressures ( $\circ$ ) of the inflections in the isothermal compression curves as a function of  $T$ , in relation to the location of the liquid spinodal of Fig. 2c. The isothermal compression curves are generated in the following way: the equilibrated system configuration obtained from the simulation at  $\rho = 0.87 \text{ g cm}^{-3}$ ,  $T = 250 \text{ K}$  is cooled for a period of 180 ps at constant  $V$  to  $200 \text{ K}$ , and then for another 50 ps to each of several different temperatures,  $T = 80 \text{ K}$ ,  $130 \text{ K}$ ,  $180 \text{ K}$  and  $235 \text{ K}$ . Each of these systems is then used as the initial (LDA ice-like) configuration for a constant- $P$ , constant- $T$  simulation, beginning at the nominal  $P$  and  $T$  resulting from the cooling procedure. During the simulation,  $P$  is increased in increments of  $25 \text{ MPa}$  every  $10 \text{ ps}$ , while  $T$  is maintained via Berendsen's method of velocity rescaling with relaxation time  $0.5 \text{ ps}$ . To check this result, we found that simulations using different starting configurations,  $P$  increments of  $50 \text{ MPa}$  every  $10 \text{ ps}$ , or  $P$  increments of  $25 \text{ MPa}$  every  $20 \text{ ps}$ , do not give significantly different results. These isothermal compressions represent non-equilibrium processes, because the systems are at too low a temperature to have time to relax to equilibrium within the stated simulation scheme. The  $T = 235 \text{ K}$  run therefore serves as a check on the deviation induced by pushing the system kinetically through an entire isotherm. As seen in this figure, there is a  $\sim 200 \text{ MPa}$  pressure gap between the  $T = 235 \text{ K}$  isothermal compression run, and the equilibrium simulation points at the same  $T$ .

(1) The low-temperature isotherms (Fig. 2b) show an intriguing trend: each exhibits an inflection in the region around a density  $\rho = 1.00 \text{ g cm}^{-3}$  which increases in strength as  $T$  is decreased. The shape and  $T$  dependence of these inflections are similar to what would occur for isotherms if a critical point were approached from above in  $T$ .

(2) The existence of a critical point  $C'$  would imply the onset of a two-phase coexistence region, so we next explore the two phases that must become distinct below  $C'$ . The structure of the system at densities just above and just below the estimated critical density  $\rho_{C'} = 1.00 \text{ g cm}^{-3}$  is reflected in Fig. 3, which shows the pair correlation function  $h(r)$  at a temperature close to the estimated critical temperature  $T_{C'}$ . Our calculations are compared with the experimentally measured  $h(r)$  for the two amorphous ice phases, high-density amorphous ice (HDA) and low-density amorphous ice (LDA)<sup>19</sup>. The similarity between

$h(r)$  for the HDA phase and our ST2 calculations just above  $\rho_{C'}$  as well as the similarity between the LDA phase and our ST2 calculations just below  $\rho_{C'}$  supports the possibility that  $C'$  is a critical point for a phase transition between a HDA-like substance and an LDA-like substance. Indeed, the fact that the two phases are both amorphous allows the possibility of a second-order transition.

(3) The phase transition from LDA to HDA ice has been observed in isothermal compression experiments<sup>7</sup> at 77 K. If  $C'$  were the critical point of the LDA-HDA ice transition in ST2 water, then simulated isothermal compressions of the ST2 system at temperatures below  $T_{C'}$  should reproduce this LDA-HDA ice transition. The isothermal compression curves for temperatures below  $T_{C'}$  are shown in Fig. 4. The analogue of the LDA-HDA transition is indicated by the sharp inflections in the isotherms, which are plausibly continuous with those generated from the liquid-state simulations. Our confidence in the reliability of conclusions drawn from the ST2 potential is enhanced by our finding that ST2 can reproduce the experimentally observed LDA-HDA transition.

(4) To confirm that the phase behaviour, reflected in Fig. 1b, is consistent with experimental results, we compare in Fig. 5 the  $T$  dependence of density,  $K_T$ , and isobaric specific heat at ambient pressure.

Before concluding, we note that our ST2 calculations lead to an overall phase diagram of the generic form displayed in Fig. 1b. The slope of the TMD line changes sign at low pressures, avoiding an intersection with the liquid spinodal, and the possibility of re-entrance of this spinodal. The HDA/LDA

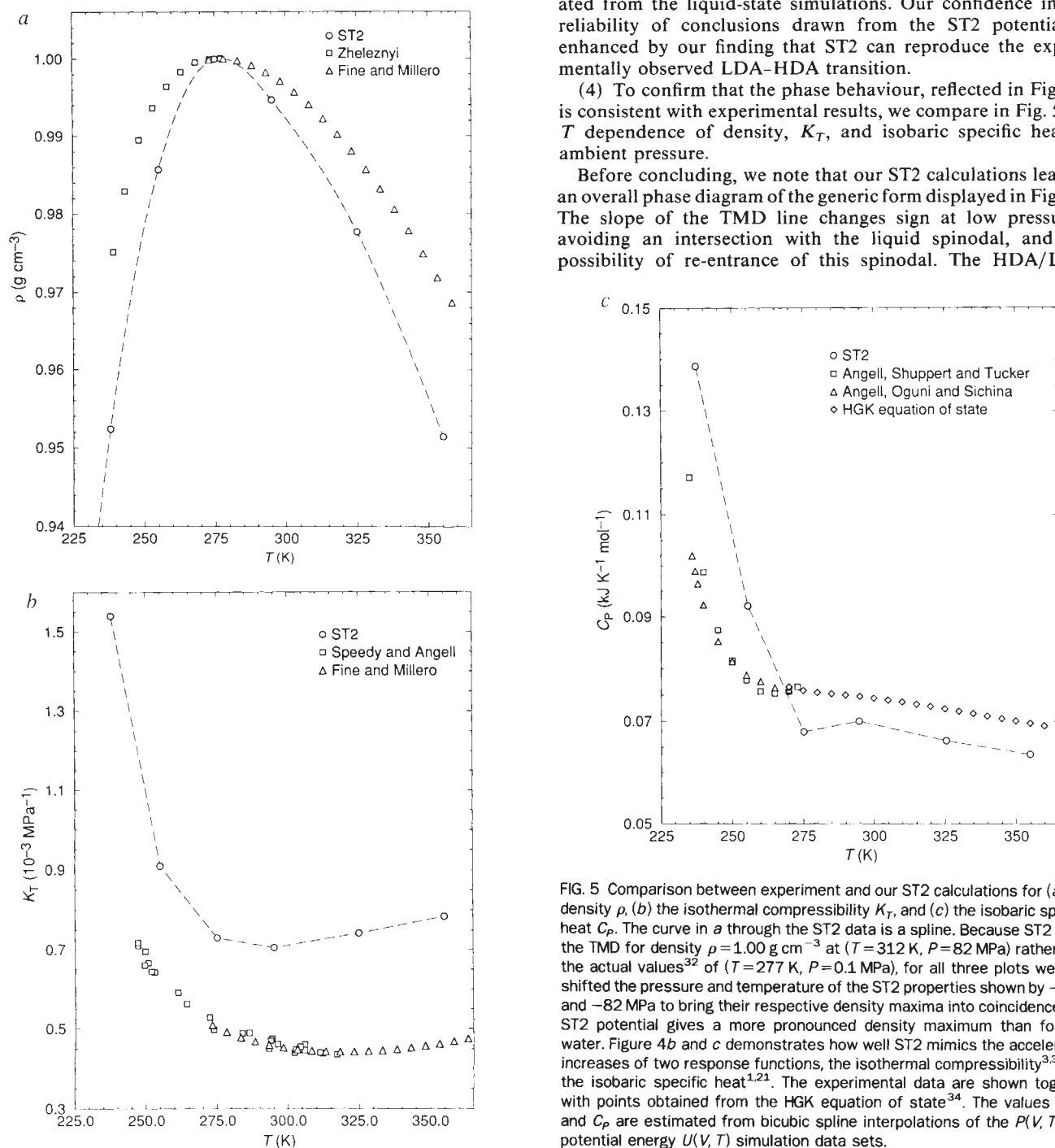


FIG. 5 Comparison between experiment and our ST2 calculations for (a) the density  $\rho$ , (b) the isothermal compressibility  $K_T$ , and (c) the isobaric specific heat  $C_p$ . The curve in *a* through the ST2 data is a spline. Because ST2 gives the TMD for density  $\rho = 1.00 \text{ g cm}^{-3}$  at  $(T = 312 \text{ K}, P = 82 \text{ MPa})$  rather than the actual values<sup>32</sup> of  $(T = 277 \text{ K}, P = 0.1 \text{ MPa})$ , for all three plots we have shifted the pressure and temperature of the ST2 properties shown by  $-35 \text{ K}$  and  $-82 \text{ MPa}$  to bring their respective density maxima into coincidence. The ST2 potential gives a more pronounced density maximum than for real water. Figure 4b and c demonstrates how well ST2 mimics the accelerating increases of two response functions, the isothermal compressibility<sup>3,33</sup> and the isobaric specific heat<sup>1,21</sup>. The experimental data are shown together with points obtained from the HGK equation of state<sup>34</sup>. The values of  $K_T$  and  $C_p$  are estimated from bicubic spline interpolations of the  $P(V, T)$  and potential energy  $U(V, T)$  simulation data sets.



coexistence line ending in  $C'$  has negative slope for the same reason that the coexistence line separating ice  $I_h$  and liquid water has negative slope: both the density and entropy of the lower-temperature phase are smaller than those of the higher-temperature phase<sup>20</sup>. There are several important implications of Fig. 1b for our understanding of amorphous ice and liquid water. First, the appearance of the critical point  $C'$  in super-cooled water and that of the density maximum are intimately related. The isotherm inflections (Fig. 2b) that eventually develop into the critical point  $C'$  represent a behaviour that simultaneously allows for the increase of  $P$  along an isochore as  $T$  is lowered, a basic signature of an anomalous liquid. Second, the behaviour that we find is consistent with suggestions<sup>21–23</sup> that the development of a random tetrahedral network as the temperature of the liquid is lowered could occur as some sort of phase transition. Third, both amorphous ice phases are incorporated into our understanding of the liquid state, a subject of recent debate<sup>24,25</sup>. Specifically, we predict that LDA ice is continuously connected to the liquid by any thermodynamic path that avoids the line of first-order transitions ending at  $C'$ . Finally, the phase behaviour proposed here may be important for understanding the properties of water in confined geometries, where behaviour similar to that of the bulk liquid is exhibited, but at comparatively higher temperatures<sup>26</sup>, thereby allowing the bulk phenomena occurring deep in the metastable region to be more readily observed. □

Received 29 July; accepted 13 October 1992.

1. Angell, C. A. in *Water: A Comprehensive Treatise* (ed. Franks, F.) Vol. 7, 1–81 (Plenum, New York, 1982).

2. Lang, E. W. & Lüdemann, H.-D. *Angew. Chem. int. Ed. Engl.* **21**, 315–329 (1982).
3. Speedy, R. J. & Angell, C. A. *J. chem. Phys.* **65**, 851–858 (1976).
4. Speedy, R. J. *J. Phys. Chem.* **86**, 982–991 (1982).
5. Speedy, R. J. *J. Phys. Chem.* **86**, 3002–3005 (1982).
6. Mishima, O., Calvert, L. D. & Whalley, E. *Nature* **310**, 393–395 (1984).
7. Mishima, O., Calvert, L. D. & Whalley, E. *Nature* **314**, 76–78 (1985).
8. Gunton, J. D., San Miguel, M. & Sahni, P. S. in *Phase Transitions and Critical Phenomena* (eds Domb, C. & Lebowitz, J. L.) 267–482 (Academic, London, 1983).
9. Compagner, A. *Physica* **72**, 115–122 (1974).
10. Debenedetti, P. G. & D'Antonio, M. C. *J. chem. Phys.* **84**, 3339–3345 (1986).
11. Debenedetti, P. G. & D'Antonio, M. C. *Am. Inst. chem. Eng. J.* **34**, 447–455 (1988).
12. Debenedetti, P. G., Raghaven, V. S. & Borick, S. S. *J. phys. Chem.* **95**, 4540–4551 (1991).
13. Green, J. L., Durben, D. J., Wolf, G. H. & Angell, C. A. *Science* **249**, 649–652 (1990).
14. Henderson, S. J. & Speedy, R. J. *J. phys. Chem.* **91**, 3062–3068 (1987).
15. Stillinger, F. H. & Rahman, A. *J. chem. Phys.* **60**, 1545–1557 (1974).
16. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. *J. chem. Phys.* **79**, 926–935 (1983).
17. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. *J. phys. Chem.* **91**, 6269–6271 (1987).
18. Strieman, L. thesis, Univ. of Dortmund (1992).
19. Bellissent-Funel, M.-C., Teixeira, J. & Bosio, L. *J. chem. Phys.* **87**, 2231–2235 (1987).
20. Whalley, E., Klug, D. D. & Handa, Y. P. *Nature* **342**, 782–783 (1989).
21. Angell, C. A., Shuppert, J. & Tucker, J. C. *J. phys. Chem.* **77**, 3092–3099 (1973).
22. Stanley, H. E. & Teixeira, J. *J. chem. Phys.* **73**, 3404–3422 (1980).
23. Sciortino, F., Poole, P., Stanley, H. E. & Havlin, S. *Phys. Rev. Lett.* **64**, 1686–1689 (1990).
24. Hallbrucker, A., Mayer, E. & Johari, G. P. *Phil. Mag.* **B60**, 179–187 (1989).
25. Speedy, R. J. *J. phys. Chem.* **96**, 2322–2325 (1992).
26. Dore, J. in *Correlations and Connectivity* (eds Stanley, H. E. & Ostrowsky, N.) 188–197 (Kluwer, Dordrecht, 1990).
27. Stillinger, F. H. & Weber, T. A. *J. chem. Phys.* **68**, 3837–3844 (1978).
28. Sciortino, F., Geiger, A. & Stanley, H. E. *Nature* **354**, 218–221 (1991).
29. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. *J. phys. Chem.* **81**, 3684–3690 (1984).
30. Chowdhury, M. R., Dore, J. C. & Wenzel, J. T. *J. non-cryst. Solids* **53**, 247–265 (1982).
31. Tse, J. S. & Klein, M. L. *Phys. Rev. Lett.* **58**, 1672–1675 (1987).
32. Zhelezni, B. V. *Russ. J. Phys. Chem.* **43**, 1311–1312 (1969).
33. Fine, R. A. & Millero, F. J. *J. chem. Phys.* **59**, 5529–5536 (1973).
34. Haar, L., Gallagher, J. S. & Kell, G. *NBS/NRC Steam Tables* (Hemisphere, Washington DC, 1985).

ACKNOWLEDGEMENTS. We thank A. Geiger, S. C. Glotzer, S. Schwarzer and especially C. A. Angell and S. Sastry for discussions and suggestions on the manuscript, and L. Strieman for sharing his preliminary results. This work was supported by the NSF, ONR and BP.

## Solar cycle length, greenhouse forcing and global climate

P. M. Kelly & T. M. L. Wigley

Climatic Research Unit, University of East Anglia, Norwich NR4 7TJ, UK

THE recent rise in global-mean surface air temperature is widely thought to be the result of increasing atmospheric concentrations of greenhouse gases<sup>1–3</sup>, but there are discrepancies between the predicted response of the atmosphere to this radiative forcing and the observed temperature changes<sup>1–5</sup>. Solar irradiance fluctuations have been proposed as a possible explanation for these discrepancies, and various solar properties (for example, radius<sup>6</sup>, smoothed sunspot number<sup>7</sup> or cycle length<sup>8</sup>) have been suggested as proxies for solar irradiance variations in the absence of direct data. Here we model the effects of a combination of greenhouse and solar-cycle-length forcing and compare the results with observed temperatures. We find that this forcing combination can explain many features of the temperature record, although the results must be interpreted cautiously; even with optimized solar forcing, most of the recent warming trend is explained by greenhouse forcing.

We use an upwelling-diffusion energy-balance climate model<sup>4,5</sup> to simulate the effects of greenhouse and solar forcing over the period 1765 to 1985. In a series of analyses, we vary the climate sensitivity and a scaling factor linking changes in solar cycle length to radiative forcing, and determine the best fit between annual modelled and observed global-mean (land-plus-marine) temperature<sup>9,10</sup> by maximizing the explained variance over the period 1861–1985 (following ref. 11).

We use two records of greenhouse forcing and allied effects. The first, 'history 1', is taken from ref. 12. The second, 'history 2', has been derived by Wigley and Raper (their 'middle' aerosol forcing case)<sup>13</sup> on the basis of a recent re-evaluation of the forcing associated with, *inter alia*, ozone depletion and sulphur dioxide emissions<sup>14</sup>. The climate sensitivity is specified by the

equilibrium global-mean warming for a change in forcing (of any origin) equivalent to a doubling of the CO<sub>2</sub> concentration,  $\Delta T_{2\times}$ , and is allowed to vary between 0.5 and 5.5 °C. This exceeds the range of uncertainty associated with cloud feedback and other processes, 1.5 to 4.5 °C, based on general circulation model calculations<sup>1</sup>. We assume a simple linear relationship between irradiance and cycle length, consistent with ref. 8. The solar forcing term is defined by  $\Delta Q_s = \beta \Delta L$ , where  $\Delta Q_s$  is the solar forcing change at the top of the troposphere,  $\Delta L$  refers to the series of cycle-length departures from the 1765–1985 mean and  $\beta$  is a solar forcing scaling factor ( $\text{W m}^{-2} \text{yr}^{-1}$ ).  $\beta$  is varied between  $-1.0$  and  $+0.5$  which implies a range of  $-3$  to  $+2 \text{ W m}^{-2}$  for  $\Delta Q_s$  ( $-1.3$  to  $+0.8\%$  irradiance change). We make use of

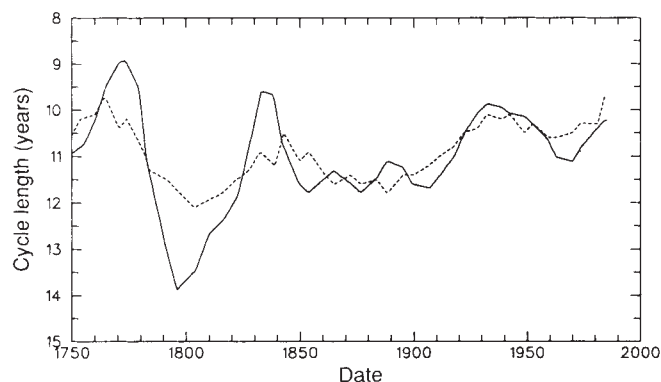


FIG. 1 The two solar cycle length records. Solid line: record derived in this paper (KW). Dashed line: record derived by Friis-Christensen and Lassen<sup>8</sup> (FCL). KW is derived from the same basic sunspot data set as FCL but the raw maximum-to-maximum and minimum-to-minimum length series were combined into a single series before applying a 7-term binomial filter. Differences between the records arise mainly because of the different bandwidths of the filters used. We have used the FCL record as published, although we found that they omitted two minor troughs around 1800 and that a different length for the final solar cycle is obtained when making use of the most recently available data.

TABLE 1 Effect of different forcings

| Forcing    |       | $\Delta T_{2x}$<br>(°C) | $\beta$<br>(W m <sup>-2</sup><br>yr <sup>-1</sup> ) | Explained variance (%) |       |       |
|------------|-------|-------------------------|-----------------------------------------------------|------------------------|-------|-------|
| Greenhouse | Solar |                         |                                                     | Greenhouse             | Solar | Total |
| History 1  | None  | 1.5                     |                                                     | 50.9                   |       | 50.9  |
| History 2  | None  | 3.7                     |                                                     | 55.7                   |       | 55.7  |
| History 1  | KW    | 0.9                     | -0.58                                               | 39.2                   | 21.6  | 60.8  |
| History 2  | KW    | 2.0                     | -0.29                                               | 41.2                   | 19.4  | 60.6  |
| History 1  | FCL   | 0.8                     | -0.59                                               | 34.5                   | 21.1  | 55.6  |
| History 2  | FCL   | 2.2                     | -0.21                                               | 44.8                   | 12.8  | 57.6  |
| None       | KW    | 25.4                    | -0.27                                               |                        | 61.6  | 61.6  |
| None       | FCL   | 12.0                    | -0.31                                               |                        | 59.5  | 59.5  |

The climate sensitivity,  $\Delta T_{2x}$ , solar forcing scaling factor,  $\beta$ , and percentage of the observed temperature variance explained for the best-fit simulations associated with the greenhouse-forcing-alone, greenhouse-plus-solar and solar-forcing-alone analyses. History 1, greenhouse forcing<sup>12</sup>; History 2, greenhouse plus aerosol forcing<sup>13</sup>; KW, present cycle-length record; FCL, Friis-Christensen and Lassen cycle-length record<sup>8</sup>.

two cycle length records, that presented by Friis-Christensen and Lassen<sup>8</sup>, FCL, and our own estimate, KW (Fig. 1).

Considerable margins of uncertainty are associated with the best-fit climate sensitivities identified in an analysis of this type<sup>3</sup>. First and foremost, we are intentionally considering a limited subset of the potential causes of recent longer-term climate change. Other mechanisms must have played a part<sup>2-5</sup> and their inclusion could alter the best-fit climate sensitivities significantly and lead to improved values for the explained variance. (Some mechanisms, such as internal climate variability<sup>3-5</sup>, provide an alternative explanation for the discrepancies between the modelled response to greenhouse forcing and observed trends.) Second, uncertainties exist in the data records used in the analysis and aspects of model formulation, although these are well documented<sup>1</sup>. These problems are unlikely to affect the relative levels of the greenhouse and solar contributions, but they could well influence their absolute levels.

We consider first greenhouse forcing alone (Table 1). For history 1, the best fit occurs when  $\Delta T_{2x} = 1.5^\circ\text{C}$ , with 50.9% of the variance in the observed temperature record accounted for. For history 2, the best fit occurs at a higher climate sensitivity,  $\Delta T_{2x} = 3.7^\circ\text{C}$ , as the global forcing is not as great. This simulation explains 55.7% of the temperature variance. These simulations capture the magnitude of the observed long-term trend in a realistic fashion, but the observed and simulated series differ noticeably on the decade-to-decade timescale (Fig. 2a and b).

We now combine greenhouse and solar forcing to determine whether a better fit between modelled and observed temperature can be obtained by including the solar component. For history 1 and the KW solar record, the best fit occurs at  $\Delta T_{2x} = 0.9^\circ\text{C}$  and  $\beta = -0.58$ . With history 2 and the KW record, the best fit is for  $\Delta T_{2x} = 2.0^\circ\text{C}$  and  $\beta = -0.29$ . The variance explained by these two simulations is 60.8% and 60.6%, respectively. The results based on the FCL record are similar, although the variance explained is slightly less (Table 1). The solar term clearly improves the agreement between the modelled and observed series (Fig. 2c-f); in particular, the marked warming around 1930 and the stable temperatures during the 1950s are better represented. If the explained variance is partitioned into greenhouse and solar components, greenhouse forcing has the strongest influence in all cases although the solar contribution is not negligible (Table 1).

Finally, we consider solar forcing alone. No best fit occurs within the specified range of  $\Delta T_{2x}$  whichever cycle length record is used. The optimal pairing of  $\beta$  and  $\Delta T_{2x}$  is found at  $\Delta T_{2x} = 25^\circ\text{C}$  for the KW record and at  $\Delta T_{2x} = 12^\circ\text{C}$  for the FCL record. The variance explained by this pair of experiments is close to 60%, and the visual correspondence between the simulated and observed temperature series is good (Fig. 2g and h). Although

these results might be considered implausible because of the extreme climate sensitivities implied, the explained variance at more reasonable climate sensitivities within the upper half of the accepted range of uncertainty, 1.5 to 4.5 °C, remains above 55%. Nevertheless, the overall credibility of the experiment with solar forcing alone must be considered low, as it is illogical to neglect greenhouse forcing given the well-established case for its existence<sup>1,2</sup>.

We now assess the plausibility of the greenhouse-plus-solar simulations in terms of their implied climate sensitivities and solar forcing ranges. The best-fit results using history 1 imply climate sensitivities below the accepted range, but they are not beyond the bounds of possibility. For history 2, the implied climate sensitivities are within the accepted range. We next compare the range of past variations in solar irradiance implied by the best-fit values of  $\beta$  with what is known from other sources. The solar and greenhouse components of the radiative forcing for the four best-fit cases are shown in Fig. 3. For the twentieth century as a whole, the variations in solar irradiance implied by the analyses are within the bounds of possibility. For earlier times, though, there are reasons for doubting certain simulations. For history 1 and the KW solar record, the implied variation in solar irradiance before 1850 is much greater than current estimates of the potential range of variability of this quantity. For

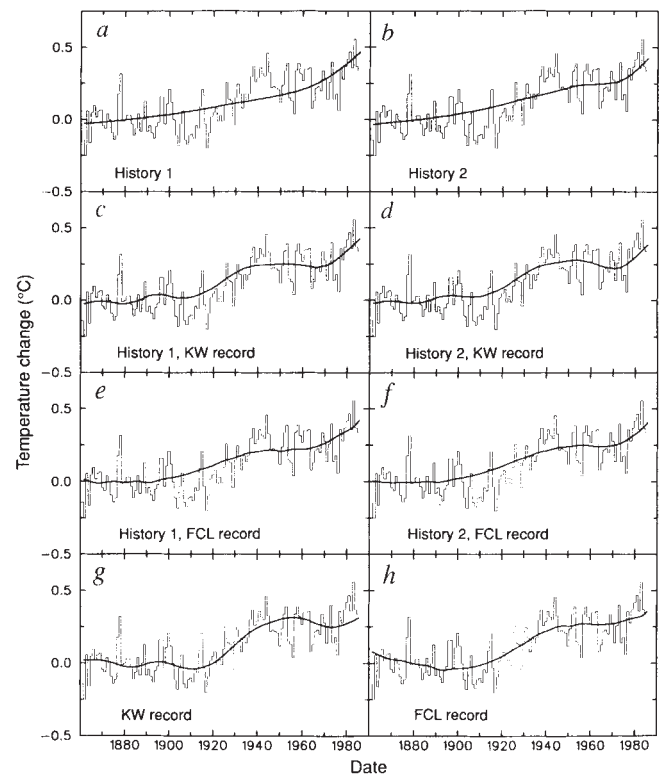


FIG. 2 Comparison of modelled (thick) and observed (thin) global-mean surface air temperature changes as departures from the 1861-1900 mean. (See Table 1 for associated CO<sub>2</sub>-doubling temperatures and solar forcing scaling factors.) We use global-mean (land-plus-marine) temperature<sup>10</sup> as it is the best available indicator for the Earth as a whole<sup>9,10,17</sup>. Ref. 8 used the Northern Hemisphere, land-based temperature record as an indicator of global change, claiming these data to be superior to the full global data set, but this is incorrect<sup>9,10</sup>. Before calculation of the explained variance, both simulated and observed temperatures have been adjusted to have the same 1861-1900 mean. We have not filtered the annual data to avoid both a subjective decision on filter characteristics and the inflation of explained variance at the expense of loss of degrees of freedom. a, Greenhouse forcing alone, history 1. b, Greenhouse forcing alone, history 2. c, Greenhouse, history 1, and solar, KW record, forcing. d, Greenhouse, history 2, and solar, KW record, forcing. e, Greenhouse, history 1, and solar, FCL record, forcing. f, Greenhouse, history 2, and solar, FCL record, forcing. g, Solar forcing alone, KW record. h, Solar forcing alone, FCL record.



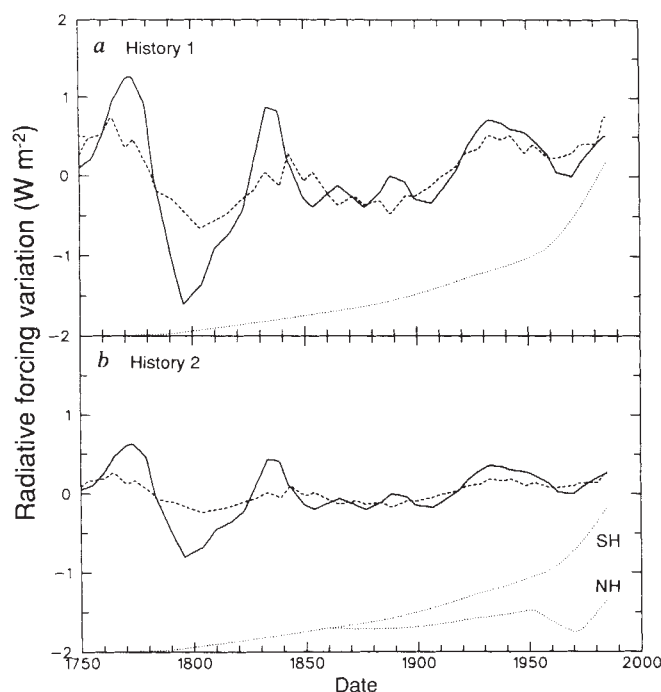


FIG. 3 Solar forcing variations ( $\text{W m}^{-2}$ ) implied by the best-fit estimates of the solar forcing scaling factor. *a*, KW (solid line) and FCL (dashed line) cycle-length records, with global history 1 greenhouse forcing (dotted, offset by  $-2 \text{ W m}^{-2}$ ). *b*, KW (solid line) and FCL (dashed line) cycle-length records, with history 2 greenhouse plus aerosol forcing for the Northern Hemisphere, NH, and Southern Hemisphere, SH (dotted, offset by  $-2 \text{ W m}^{-2}$ ). The greenhouse plus aerosol forcing in this case is hemisphere-specific because of the differential effects of the sulphate aerosol contribution.

example, the solar forcing increases by about  $2.5 \text{ W m}^{-2}$  from 1795 to 1834, an increase of more than 1% in irradiance. Such a change is unlikely. It is well beyond both the departure necessary to account for the Little Ice Age<sup>15</sup> and, more important, beyond the range of potential departures estimated from astronomical data<sup>16</sup>. In two of the other cases (history 2, KW record; history 1, FCL record), the implied irradiance variations about the year 1800 border on inconsistency with the independent evidence cited above. Only in the case of history 2 and the FCL record are the implied irradiance changes compatible with that evidence. In this case, the maximum change over a 40-year period is  $-0.5 \text{ W m}^{-2}$  ( $-0.21\%$ ) from 1765 to 1804. A nonlinear  $\Delta Q_s - \Delta L$  relationship may reduce the range of implied solar changes before 1900, but there is no evidence available at present to support a more complex relationship than assumed here.

In summary, our results provide circumstantial support for the hypothesized link between solar cycle length and irradiance. Even in the optimal cases considered here, however, the solar contribution to recent global-mean temperature change is much less than that due to increasing greenhouse gas concentrations and other anthropogenic effects. The value of cycle length data is limited because our results show considerable sensitivity to the filtering method used to construct the record. This implies large uncertainties in the cycle-length-based record of solar irradiance changes over and above those associated with the optimization procedure used to deduce  $\beta$ . The physical mechanism underlying the proposed link between solar output and cycle length must be better understood before cycle length can be used with confidence as a proxy for irradiance changes.

**Note added in proof:** The accompanying paper by Schlesinger and Ramankutty, which was shown to us after we had submitted this paper, also considers the relative importance of greenhouse, sulphate aerosol and solar forcing. There are minor differences in the results which arise mainly because we use the latest IPCC temperature data<sup>10</sup> and include stratospheric ozone depletion

feedback in history 2 (ref. 13). In addition, the model formulation and parameter values used<sup>13</sup> differ slightly between the two analyses. □

Received 7 July; accepted 19 October 1992.

- Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, 1990).
- Hansen, J. E. & Lacis, A. A. *Nature* **346**, 713–719 (1990).
- Wigley, T. M. L. & Raper, S. C. B. in *Climate Change: Science, Impacts and Policy* (eds Jäger, J. & Ferguson, H. L.) 231–242 (Cambridge Univ. Press, 1991).
- Wigley, T. M. L. & Raper, S. C. B. *Nature* **330**, 127–131 (1987).
- Wigley, T. M. L. & Raper, S. C. B. *Nature* **344**, 324–327 (1990).
- Gilliland, R. L. *Clim. Change* **4**, 111–131 (1982).
- Reid, G. C. *Nature* **329**, 142–143 (1987).
- Friis-Christensen, E. & Lassen, K. *Science* **254**, 698–700 (1991).
- Folland, C. K., Karl, T. R. & Vinnikov, K. Ya. in *Climate Change: The IPCC Scientific Assessment* (eds Houghton, J. T., Jenkins, G. J. & Ephraums, J. J.) 195–238 (Cambridge Univ. Press, 1990).
- Folland, C. K. et al. in *Climate Change 1992: Supplementary Report to the IPCC Scientific Assessment* (eds Houghton, J. T., Callander, B. A. & Varney, S. K.) 135–170 (Cambridge Univ. Press, 1992).
- Kelly, P. M. & Wigley, T. M. L. *Nature* **347**, 460–462 (1990).
- Shine, K. P., Derwent, R. G., Wuebbles, D. J. & Morcrette, J.-J. in *Climate Change: The IPCC Scientific Assessment* (eds Houghton, J. T., Jenkins, G. J., & Ephraums, J. J.) 41–68 (Cambridge Univ. Press, 1990).
- Wigley, T. M. L. & Raper, S. C. B. *Nature* **357**, 293–300 (1992).
- Isaksen, I. S. A. et al. in *Climate Change 1992: Supplementary Report to the IPCC Scientific Assessment* (eds Houghton, J. T., Callander, B. A. & Varney, S. K.) 47–67 (Cambridge Univ. Press, 1992).
- Wigley, T. M. L. & Kelly, P. M. *Phil. Trans. R. Soc. A* **330**, 547–560 (1990).
- Baliunas, S. & Jastrow, R. *Nature* **348**, 520–523 (1990).
- Jones, P. D., Wigley, T. M. L. & Farmer, G. in *Greenhouse-Gas-Induced Climatic Change: A Critical Appraisal of Simulations and Observations* (ed. Schlesinger, M. E.) 153–172 (Elsevier, 1991).

ACKNOWLEDGEMENTS. This analysis was supported by the US Department of Energy, Atmospheric and Climate Research Division.

## Implications for global warming of intercycle solar irradiance variations

Michael E. Schlesinger & Navin Ramankutty

Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, 105 South Gregory Avenue, Urbana, Illinois 61801, USA

FOLLOWING earlier studies<sup>1–6</sup>, attention has recently been directed again to the possibility that long-term solar irradiance variations, rather than increased greenhouse gas concentrations, have been the dominant cause of the observed rise in global-mean surface temperature from the mid-nineteenth century to the present. Friis-Christensen and Lassen<sup>7</sup> report a high correlation (0.95; ref. 8) between the variable period of the '11-year' sunspot cycle and the mean Northern Hemisphere land surface temperature from 1865 to 1985. The Marshall Institute report<sup>9</sup> concludes that '... the sun has been the controlling influence on climate in the last 100 years, with the greenhouse effect playing a smaller role.' Here we explore the implication that such putative solar irradiance variations would have for global warming. Our results provide strong circumstantial evidence that there have been intercycle variations in solar irradiance which have contributed to the observed temperature changes since 1856. However, we find that since the nineteenth century, greenhouse gases, not solar irradiance variations, have been the dominant contributor to the observed temperature changes.

We use a hemispherically resolving version<sup>10</sup> of our simple climate-ocean model which we used for IPCC/90 (ref. 11) and other studies<sup>12–14</sup>. The model calculates the change in hemispheric-mean surface air temperature ( $\Delta T_a$ ) and ocean temperatures as a function of depth resulting from a change in hemispheric-mean net radiation at the tropopause due to external cause(s). The temperature change  $\Delta T_a$  generates an opposing change in tropopause radiative flux,  $-\lambda \Delta T_a$ , where  $\lambda$  is a prescribed sensitivity parameter which implicitly includes the effects of changes in feedback quantities—for example, water vapour, clouds, and snow and ice—as well as in temperature. We assume

that the same  $\lambda$  is valid for each of the radiative forcings studied here. Earlier studies indicate that the error associated with this assumption is small, as simple climate models reproduce the response of more-complex climate models<sup>15,16</sup> and these show similar responses to CO<sub>2</sub> and solar-constant changes<sup>17</sup>. For convenience we write  $\lambda = \Delta F_{2\times} / \Delta T_{2\times}$ , where  $\Delta T_{2\times}$  is a prescribed equilibrium value of  $\Delta T_a$  resulting from a radiative forcing equivalent to that for doubling the CO<sub>2</sub> concentration,  $\Delta F_{2\times} = 4.4 \text{ W m}^{-2}$  (ref. 18). The latter is based on radiative-transfer-model calculations for standard and doubled CO<sub>2</sub> (such as ICRCCM<sup>19</sup>) which do not include any solar-variability effects.

The model is forced by time-dependent tropopause radiative changes due to increasing greenhouse gases (GHGs), intercycle solar irradiance variations and increasing anthropogenic sulphate aerosols (ASAs). The forcing for GHGs is

$$\Delta F_{\text{GHG}}(t) = 4.4 \frac{\ln [C(t)/C(1765)]}{\ln 2}$$

where  $C(t)$  is the historical annual-mean equivalent-carbon-dioxide concentration from 1765 to 1990 determined from IPCC/90 (ref. 18). The intercycle solar-irradiance forcing is taken as  $\Delta F_{\text{Sun}}(t) = \beta[P_0 - P(t)]$ , where  $P(t)$  is the length of the '11-year' sunspot cycle given by Gleissberg<sup>20</sup> for June 1638 to December 1744 and Friis-Christensen and Lassen<sup>7</sup> for January 1745 to June 1985,  $P_0 = 11.018 \text{ yr}$  is the average cycle period over June 1638 to June 1985, and  $\beta$  represents the putative variation of solar irradiance with respect to  $P$ . We ignore here intra-cycle solar-irradiance forcing<sup>21</sup> because studies have shown its temperature response to be negligibly small<sup>22-25</sup>. The forcing due to increasing ASAs is taken to be

$$\Delta F_{\text{SO}_4}(t) = \Delta F_{\text{SO}_4}(1978) \frac{Q_{\text{SO}_2}(t)}{Q_{\text{SO}_2}(1978)}$$

where  $Q_{\text{SO}_2}(t)$  is an estimated emission rate of sulphur in the form of SO<sub>2</sub> from 1860 through 1985 and  $\Delta F_{\text{SO}_4}(1978)$  is the radiative forcing of ASAs in 1978, this being the year for which Charlson *et al.*<sup>26</sup> made detailed radiative-transfer calculations for the part of  $\Delta F_{\text{SO}_4}$  contributed by the clear atmosphere. But

because both the clear and cloudy parts of  $\Delta F_{\text{SO}_4}$  are uncertain<sup>10,26</sup>, we prescribe  $\Delta F_{\text{SO}_4}$  from zero to  $-1.7 \text{ W m}^{-2}$  on the basis of an uncertainty analysis of  $\Delta F_{\text{SO}_4}$  for the clear atmosphere combined with a rough estimate of  $\Delta F_{\text{SO}_4}$  for the cloudy atmosphere<sup>10</sup>.

We estimate  $\Delta T_{2\times}$  for three cases: (1) GHG without and with ASA; (2) Sun without and with ASA; and (3) GHG+Sun without and with ASA. Case (2) cannot occur by itself, however, because, for any climate sensitivity  $\lambda$ , the radiative forcing due to the observed increase in GHGs will induce a warming which cannot be ignored. Consequently, case (2) is included here only as an informative comparison with case (1), a case which can exist alone if  $\beta = 0$ .

For GHG alone and GHG with ASA, the model was run for many values of  $\Delta T_{2\times}$ , each beginning at  $t_0 = \text{June 1638}$  with  $\Delta T_m(t_0) = 0$ , where  $\Delta T_m$  is the global average of the model's changes in mixed-layer temperature over ocean and air temperature over land, and extending through June 1985. The model results were then compared with observations of the departure of global-mean, annual-mean temperature from its 1951–80 average,  $\delta T_0(t)$ . This comparison was necessary because only these temperature departures have been compiled, not the actual temperatures themselves (ref. 27, updated and extended by C. Folland, D. Parker and R. Hackett, Hadley Centre, UK Meteorological Office). For each  $\Delta T_{2\times}$ , the root-mean-square error (RMSE) of model-simulated annual-mean temperature departure,  $\delta T_m(t) = \Delta T_m(t) + \delta T_{\text{PI}}$ , to  $\delta T_0(t)$  was calculated from January 1856 through June 1985, with the pre-industrial average temperature anomaly,  $\delta T_{\text{PI}}$ , determined such that the RMSE was minimized<sup>10</sup>. The  $\Delta T_{2\times}$  giving the smallest RMSE was then chosen. The same procedure was performed for Sun alone and Sun with ASA, and for GHG plus Sun, both without and with ASA, except that both  $\Delta T_{2\times}$  and  $\beta$  were determined to minimize the RMSE. We have also run the model from 1745 to 1985 without the Gleissberg data<sup>20</sup> and obtained virtually identical results.

Figure 1 presents plots of  $\Delta T_{2\times}$ ,  $\beta$ , the unexplained standard deviation (USD: standard deviation of  $\delta T_0(t) - \delta T_m(t)$  as a percentage of the standard deviation of  $\delta T_0(t)$ ), and the residual

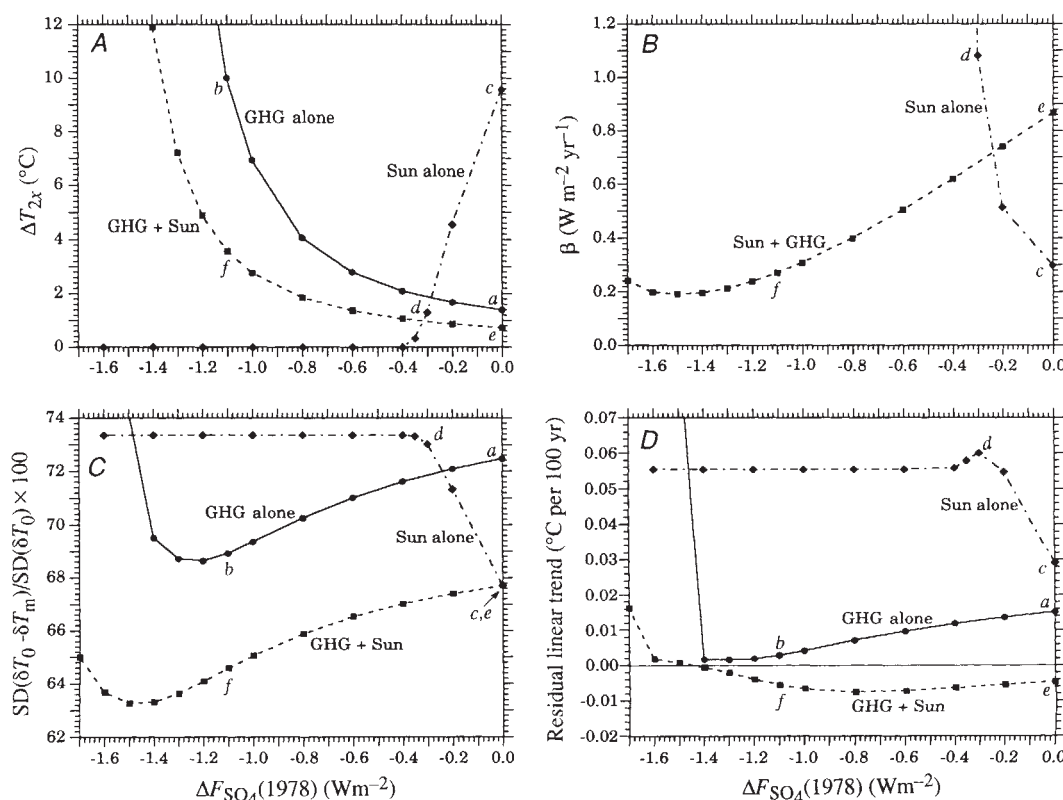


FIG. 1 Variation with ASA forcing,  $\Delta F_{\text{SO}_4}(1978)$ , of (A) climate sensitivity,  $\Delta T_{2\times}$ ; (B) intercycle solar-irradiance parameter,  $\beta$ ; (C) unexplained standard deviation, USD; and (D) residual linear trend, RLT.



linear trend (RLT: trend of  $\delta T_0(t) - \delta T_m(t)$ ) against  $\Delta F_{\text{SO}_4}(1978)$ . For  $\Delta F_{\text{SO}_4}(1978) = 0$ , Fig. 1 shows that adding solar forcing with  $\beta = 0.867 \text{ W m}^{-2} \text{ yr}^{-1}$  to GHG forcing (going from point *a* to *e*) reduces  $\Delta T_{2x}$  from  $1.4^\circ\text{C}$  to  $0.7^\circ\text{C}$ , USD from 72.5% to 67.7%, and RLT from  $0.015^\circ\text{C}$  per century to  $-0.005^\circ\text{C}$  per century. Removing GHG forcing to leave Sun alone (*e* to *c*) reduces  $\beta$  to  $0.300 \text{ W m}^{-2} \text{ yr}^{-1}$ , increases  $\Delta T_{2x}$  to  $9.6^\circ\text{C}$ , and increases RLT to  $0.029^\circ\text{C}$  per century. As negative ASA forcing is added for GHG alone and GHG plus Sun,  $\Delta T_{2x}$  increases nonlinearly<sup>10</sup>, USD decreases to minima at  $\Delta F_{\text{SO}_4}(1978) = -1.2$  and  $-1.5 \text{ W m}^{-2}$ , respectively, and the magnitude of RLT decreases to minima at  $\Delta F_{\text{SO}_4}(1978) = -1.3$  and  $-1.4 \text{ W m}^{-2}$ , respectively. In marked contrast, decreasing  $\Delta F_{\text{SO}_4}(1978)$  for Sun alone causes  $\Delta T_{2x}$  to decrease toward zero and  $\beta$  to rapidly increase such that both USD and RLT increase to constant values for  $\Delta F_{\text{SO}_4}(1978) \leq -0.4 \text{ W m}^{-2}$ . The ratio of  $\Delta T_{2x}$  for GHG plus Sun to  $\Delta T_{2x}$  for GHG alone decreases with decreasing  $\Delta F_{\text{SO}_4}(1978)$  from 0.52 for  $\Delta F_{\text{SO}_4}(1978) = 0$ .

Figure 2 illustrates the simulated temperature anomalies,  $\delta T_m$ , in comparison with the observed temperature anomaly,  $\delta T_0$ , for points *a* to *f* in Fig. 1. Points *a*, *c* and *e* have zero sulphate forcing, points *b* and *f* have  $\Delta F_{\text{SO}_4}(1978) = -1.1 \text{ W m}^{-2}$ , the central estimate of the previously mentioned uncertainty analysis<sup>10</sup>, and point *d* is an example near  $\Delta F_{\text{SO}_4}(1978) = -0.4 \text{ W m}^{-2}$  where  $\Delta T_{2x} \approx 0$  for Sun alone. For GHG alone and GHG with  $\Delta F_{\text{SO}_4}(1978) = -1.1 \text{ W m}^{-2}$  (Fig. 2A),  $\delta T_m$  respectively equals  $\delta T_{\text{PI}} = -0.34^\circ\text{C}$  and  $-0.53^\circ\text{C}$  until 1765 when GHG forcing begins. Thereafter,  $\delta T_m$  monotonically increases to 1985 when the global warming,  $\Delta T_m(t) = \delta T_m(t) - \delta T_{\text{PI}}$ , is  $0.52^\circ\text{C}$  and  $0.75^\circ\text{C}$ , respectively. In contrast, the  $\delta T_m$  values for Sun alone

and Sun with  $\Delta F_{\text{SO}_4}(1978) = -0.3 \text{ W m}^{-2}$  (Fig. 2B) oscillate about  $\delta T_{\text{PI}} = -0.13^\circ\text{C}$ , with maxima in the eighteenth, nineteenth and twentieth centuries. For Sun alone, the eighteenth- and twentieth-century maxima are comparable, whereas for Sun plus ASA the eighteenth-century maximum is larger. For GHG plus Sun (Fig. 2C),  $\delta T_m$  oscillates about  $\delta T_{\text{PI}} = -0.26^\circ\text{C}$  with no noticeable trend until about 1900. By 1922,  $\delta T_m$  exceeds its eighteenth-century maximum, continues to increase until 1944, remains nearly constant until 1966, and then increases to  $\Delta T_m(1985) = 0.42^\circ\text{C}$ . For GHG plus Sun with  $\Delta F_{\text{SO}_4}(1978) = -1.1 \text{ W m}^{-2}$ ,  $\delta T_m$  oscillates about  $\delta T_{\text{PI}} = -0.35^\circ\text{C}$  until 1911 when  $\delta T_m$  exceeds its nineteenth-century maximum. Thereafter,  $\delta T_m$  increases to 1946, remains nearly constant until 1971, and then increases to  $\Delta T_m(1985) = 0.55^\circ\text{C}$ .

The contributions of the individual forcing components to  $\Delta T_m(t)$  are presented in Fig. 3. For GHG plus Sun (Fig. 3a) the temperature change due to GHG permanently exceeds the magnitude of the temperature change due to the Sun after 1894, and their difference increases to a maximum of  $0.19^\circ\text{C}$  in 1982. The same is true for GHG plus Sun with  $\Delta F_{\text{SO}_4}(1978) = -1.1 \text{ W m}^{-2}$  (Fig. 3b), but beginning in 1827 and increasing to a maximum difference of  $0.93^\circ\text{C}$  in 1985. These results do not support the conclusion of the Marshall Institute report<sup>9</sup>, which we quoted in our opening paragraph.

Figure 4 presents the intercycle solar-irradiance anomaly,  $\Delta S = \Delta F_{\text{Sun}}[4/(1 - \alpha_p)]$ , with  $\alpha_p = 0.3$  the planetary albedo, for the two cases of Fig. 3, both in  $\text{W m}^{-2}$  and as a percentage of the mean solar irradiance measured from 1980 to 1989 by the Solar Maximum Satellite (SMS),  $S_0 = 1,367.5 \text{ W m}^{-2}$  (ref. 21, and R. C. Willson, personal communication). The intra-cycle solar-irradiance variation given by the standard deviation of the SMS observations,  $\sigma(\text{intra}) = 0.6 \text{ W m}^{-2}$  (0.044% of  $S_0$ ; R. C.

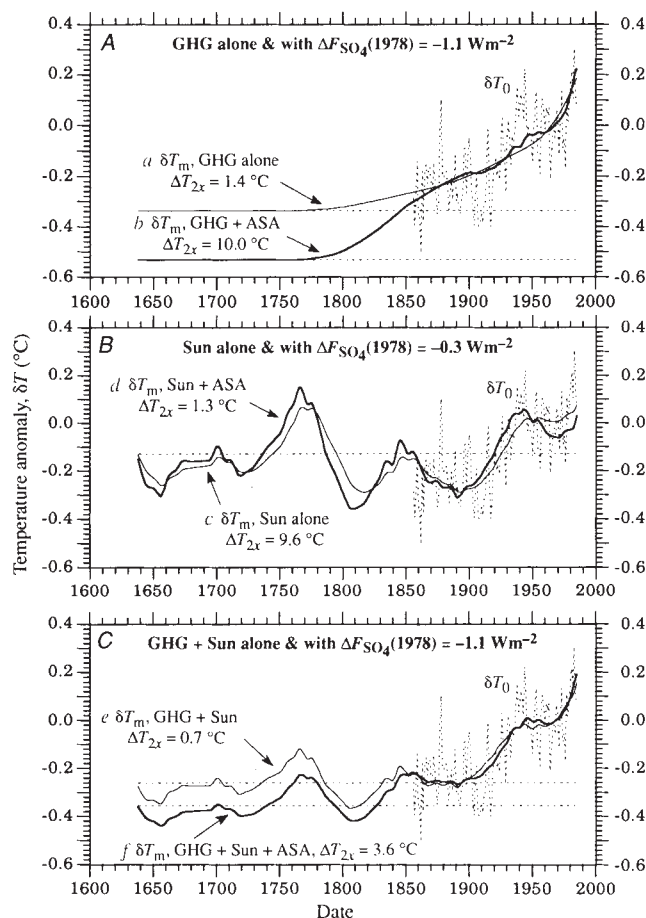


FIG. 2 Comparison of the simulated and observed temperature anomalies,  $\delta T_m$  and  $\delta T_0$ , for points *a*–*f* in Fig. 1. Dashed lines show the pre-industrial temperature anomaly,  $\delta T_{\text{PI}}$ .

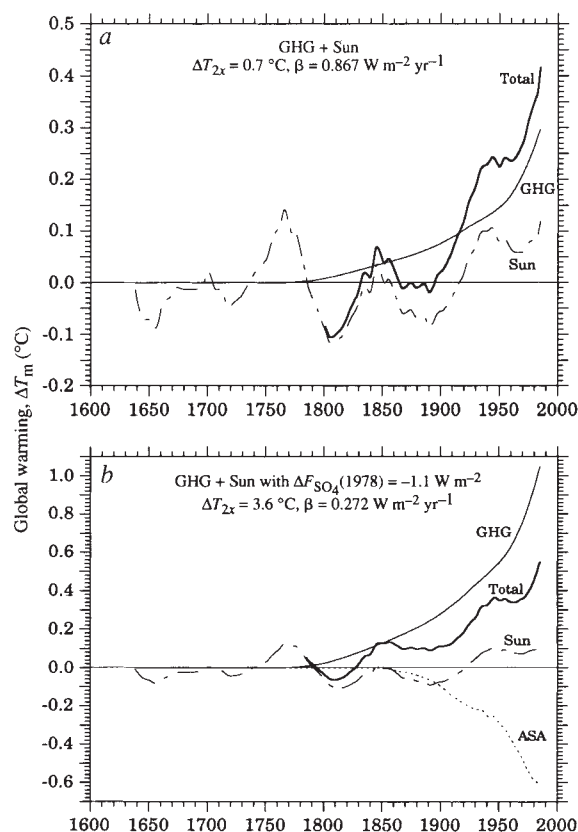


FIG. 3 Contributions of GHGs, intercycle solar variations and ASAs to global warming,  $\Delta T_m$ . The heavy line for the total  $\Delta T_m$  ('Total') is not shown before 1801 in *a* and 1785 in *b* so that solar contribution (Sun) can be seen.

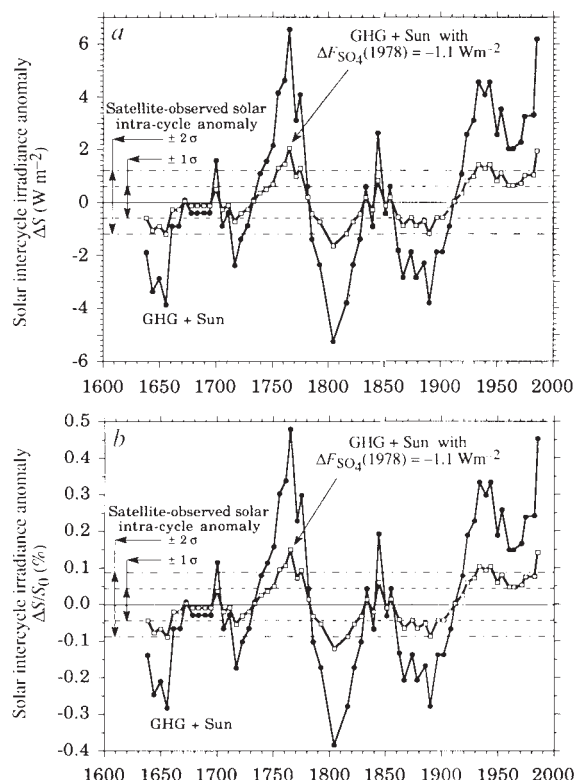


FIG. 4 Intercycle solar-irradiance variation for the two cases of Fig. 3,  $\Delta S$  (a) and  $\Delta S/S_0$  (b), where  $S_0 = 1367.5 \text{ W m}^{-2}$  is the mean solar irradiance measured by the Solar Maximum Satellite (SMS) from 1980 to 1989. The values of  $\pm\sigma$  and  $\pm 2\sigma$  for the SMS observations,  $\sigma(\text{intra}) = 0.6 \text{ W m}^{-2}$  (0.044% of  $S_0$ ), are also shown.

Willson, personal communication), is also shown for comparison. For GHG plus Sun with  $\Delta F_{\text{SO}_4}(1978) = -1.1 \text{ W m}^{-2}$ ,  $\Delta S$  ranges from  $-1.6 \text{ W m}^{-2}$  to  $2.1 \text{ W m}^{-2}$  ( $-0.12\%$  to  $0.15\%$ ), with a standard deviation  $\sigma(\text{inter}) = 0.85 \text{ W m}^{-2}$  (0.062%). This intercycle variation is within the "several tenths of a per cent" estimated by Baliunas and Jastrow<sup>28</sup> from observations of the visual-brightness and magnetic-activity variations of 74 solar-type stars. For GHG plus Sun without ASA,  $\Delta S$  ranges from  $-5.3 \text{ W m}^{-2}$  to  $6.5 \text{ W m}^{-2}$  ( $-0.38\%$  to  $0.48\%$ ), with  $\sigma(\text{inter}) = 2.71 \text{ W m}^{-2}$  (0.198%). This intercycle variation is within the 0.1–3% variation found by Lockwood *et al.*<sup>29</sup> for 25 solar-type stars. Consequently, our results constitute strong evidence, albeit circumstantial, that there have been variations in solar irradiance which have contributed to the observed temperature changes since 1856.

It can be argued that past intercycle solar-irradiance variations may have been larger than those used herein based on estimating  $\beta$  to minimize the RMSE between  $\delta T_m$  and  $\delta T_0$ . Although it is likely that significantly larger irradiance variations would have been detected by past ground-based observations<sup>30</sup>, conclusive evidence must await future satellite observations spanning several solar cycles. Regrettably, this evidence will be acquired too late to influence contemporary greenhouse policy. Nevertheless, our present evidence indicates that GHGs will continue to dominate global warming, but that inclusion of the Sun's past influence reduces the estimate of  $\Delta T_{2\times}$  by at least 48%. This does not necessarily mean that  $\Delta T_{2\times}$  is small because, as shown here, the uncertain negative radiative forcing by ASAs (and biomass burning<sup>31</sup>) can increase  $\Delta T_{2\times}$  significantly. Clearly, it is imperative to reduce this uncertainty.

The accompanying paper<sup>32</sup> also considers the relationship between solar forcing, greenhouse forcing and climate change. □

Received 27 July; accepted 19 October 1992.

- Hansen, J. *et al.* *Science* **213**, 957–966 (1981).
- Gilliland, R. L. *Clim. Change* **4**, 111–131 (1982).
- Gilliland, R. L. & Schneider, S. H. *Nature* **310**, 38–41 (1984).
- Reid, G. C. *Nature* **329**, 142–143 (1987).
- Seitz, F., Jastrow, R. & Nierenberg, W. A. *Scientific Perspectives on the Greenhouse Problem* (George C. Marshall Institute, Washington DC, 1989).
- Reid, G. C. *J. geophys. Res.* **96**, 2835–2844 (1991).
- Friis-Christensen, E. & Lassen, K. *Science* **254**, 698–700 (1991).
- Kerr, R. A. *Science* **254**, 652–653 (1991).
- Jastrow, R., Nierenberg, W. & Seitz, F. *Global Warming Update: Recent Scientific Findings* (George C. Marshall Institute, Washington DC, 1992).
- Schlesinger, M. E., Jiang, X. & Charlson, R. J. in *Climate Change and Energy Policy. Proc. int. Conf. Global Climate Change: Its Mitigation Through Improved Production and Use of Energy* (eds Rosen, L. & Glasser, R.) 75–108 (American Institute of Physics, New York, 1992).
- Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, 1990).
- Schlesinger, M. E. & Jiang, X. *Nature* **350**, 219–221 (1991).
- Hammitt, J. K., Lempert, R. J. & Schlesinger, M. E. *Nature* **357**, 315–318 (1992).
- Schlesinger, M. E. *Eos* **73**, 325–327 (1992).
- Schlesinger, M. E. in *Understanding Climate Change* (eds Berger, A., Dickinson, R. E. & Kidson, J. W.), *Geophys. Monogr.* **52**, 177–187 (American Geophysical Union, Washington DC 1989).
- Schlesinger, M. E. & Jiang, X. *J. Clim.* **3**, 1297–1315 (1990).
- Manabe, S. & Wetherald, R. T. *J. Atmos. Sci.* **37**, 99–118 (1980).
- Shine, K. P., Derwent, R. G., Wuebbles, D. J. & Morcrette J.-J. in *Climate Change: The IPCC Scientific Assessment* (eds Houghton, J. T., Jenkins, G. J. & Ephraums, J. J.) 41–68 (Cambridge Univ. Press, 1990).
- Luther, F. M. *et al.* *Bull. Am. Met. Soc.* **69**, 40–48 (1988).
- Gleissberg, W. *Terrest. Magnet. Atmos. Electr.* **49**, 243–244 (1944).
- Willson, R. C. & Hudson, H. S. *Nature* **351**, 42–44 (1991).
- Hoffert, M. I., Frei, A. & Narayanan, V. K. *Clim. Change* **13**, 267–285 (1988).
- Hoffert, M. I. in *Greenhouse-Gas-Induced Climatic Change: A Critical Appraisal of Simulations and Observations* (ed. Schlesinger, M. E.) 413–428 (Elsevier, Amsterdam, 1991).
- Wigley, T. M. L. & Raper, S. C. B. *Geophys. Res. Lett.* **17**, 2169–2172 (1990).
- Kelly, P. M. & Wigley, T. M. L. *Nature* **347**, 460–462 (1990).
- Charlson, R. J., Langner, J., Rodhe, H., Leovy, C. B. & Warren, S. G. *Tellus* **43**, 152–163 (1991).
- Folland, C. K., Karl, T. R. & Vinnikov, K. Y. in *Climate Change: The IPCC Scientific Assessment* (eds Houghton, J. T., Jenkins, G. J. & Ephraums, J. J.) 195–238 (Cambridge Univ. Press, 1990).
- Baliunas, S. & Jastrow, R. *Nature* **348**, 520–523 (1990).
- Lockwood, G. W. *et al.* *Nature* (in the press).
- Fröhlich, C. in *The Solar Output and its Variation* (ed. White, O. R.) 93–109 (Colorado Associated Univ. Press, Boulder, 1977).
- Penner, J. E., Dickinson, R. E. & O'Neill, C. A. *Science* **256**, 1432–1433 (1992).
- Kelly, P. M. & Wigley, T. M. L. *Nature* **360**, 328–330 (1992).

ACKNOWLEDGEMENTS. We thank C. Folland, D. Parker and R. Hackett for the IPCC monthly surface temperature anomaly data; J. McKinnon for the monthly sunspot record; R. Willson for the mean and standard deviation of the solar constant observed by the Solar Maximum Satellite; and A. Ghanem for assisting in the digitization of the solar-cycle length data. This research was supported by the US NSF and the US Department of Energy, Carbon Dioxide Research Program, Office of Health and Environmental Research.

## Proterozoic depletion of the lithosphere recorded in mantle xenoliths from Inner Mongolia

F.-L. Deng & J. D. Macdougall

Scripps Institution of Oceanography, La Jolla, California 92093-0220, USA

MANTLE xenoliths entrained in alkali basalt provide direct samples from a cross-section of the lithosphere. In most cases, the chemistry and isotope systematics of these objects have suggested a complex history of chemical change by processes such as metasomatism and partial melting (see, for example, ref. 1). Here, however, we present evidence from a suite of xenoliths from alkali basalts in the Chinese province of Inner Mongolia, suggesting that the lithosphere in this region has remained stable and closed to chemical modification since experiencing partial melting in the Proterozoic. Clinopyroxene separates from these xenoliths yield an approximate isochron in the  $^{143}\text{Nd}/^{144}\text{Nd}$  versus  $^{147}\text{Sm}/^{144}\text{Nd}$  diagram, suggesting an episode of depletion  $\sim 1.6$  Gyr ago. Although other explanations for this correlation, such as mixing, cannot be ruled out definitively, further evidence for the long-term undisturbed nature of the lithosphere in this region comes from neodymium model ages of the clinopyroxenes. In particular, one highly depleted sample with very high  $^{147}\text{Sm}/^{144}\text{Nd}$  provides a model age of 1.6 Gyr, consistent with the isochron result.

Alkali and tholeiitic basalts of Pleistocene age, the former often carrying ultramafic xenoliths, occur over wide areas of the Chinese province of Inner Mongolia, and also extend into



TABLE 1 Analytical data for xenolith minerals and host rocks

| Sample                | $^{143}\text{Nd}/^{144}\text{Nd}$ | $^{87}\text{Sr}/^{86}\text{Sr}$ | Sm<br>(p.p.m.) | Nd<br>(p.p.m.) | Rb<br>(p.p.m.) | Sr<br>(p.p.m.) | Equilibration<br>temperature (°C) |
|-----------------------|-----------------------------------|---------------------------------|----------------|----------------|----------------|----------------|-----------------------------------|
| <b>Clinopyroxenes</b> |                                   |                                 |                |                |                |                |                                   |
| IM8801-2              | 0.513074 ± 14                     | 0.702997 ± 26                   | 1.547          | 3.909          | 0.0060         | 62.5           | 854                               |
| IM8801-5              | 0.513266 ± 16                     | 0.702803 ± 26                   | 1.136          | 2.560          | 0.0059         | 42.9           | 874                               |
| IM8801-14A            | 0.513160 ± 18                     | 0.702634 ± 26                   | 1.580          | 3.930          | —              | 66.3           | 807                               |
| IM8808-5              | 0.512542 ± 14                     | 0.704863 ± 26                   | 2.008          | 6.565          | —              | 87.6           | 873                               |
| IM8808-10A            | 0.513265 ± 20                     | 0.702586 ± 26                   | 1.485          | 3.457          | 0.0583         | 47.3           | 863                               |
| IM8810-1A             | 0.512859 ± 14                     | 0.703289 ± 26                   | 1.734          | 4.655          | 0.0232         | 64.9           | 832                               |
| IM8814-1              | 0.513398 ± 14                     | 0.702227 ± 26                   | 1.630          | 3.746          | 0.0397         | 59.7           | 881                               |
| IM8814-2              | 0.513330 ± 14                     | 0.703464 ± 26                   | 0.3091         | 0.666          | 0.0012         | 13.8           | 961                               |
| IM8815-5B             | 0.512930 ± 14                     | 0.703770 ± 26                   | 1.289          | 3.171          | 0.0037         | 64.4           | 921                               |
| IM8815-7AB            | 0.513436 ± 16                     | 0.702709 ± 26                   | 0.9763         | 2.029          | 0.0064         | 27.2           | 965                               |
| IM8815-7D-1           | 0.515471 ± 15                     | 0.703824 ± 26                   | 0.0906         | 0.119          | —              | 4.88           | 824                               |
| IM8815-7D-2           | 0.515364 ± 14                     | 0.703758 ± 26                   | 0.0916         | 0.123          | —              | 4.91           | 824                               |
| IM8816-25             | 0.513319 ± 14                     | 0.702619 ± 26                   | 0.6006         | 1.589          | 0.0017         | 32.3           | 892                               |
| IM8817-5              | 0.512730 ± 14                     | 0.704987 ± 26                   | 1.611          | 4.198          | —              | 55.9           | 909                               |
| <b>Orthopyroxenes</b> |                                   |                                 |                |                |                |                |                                   |
| IM8801-2              | 0.513081 ± 27                     | 0.703152 ± 26                   | 0.00887        | 0.01407        | —              | 0.14           |                                   |
| IM8808-5              | 0.512584 ± 14                     | 0.704727 ± 26                   | 0.01761        | 0.04195        | 0.0207         | 0.36           |                                   |
| <b>Host basalt</b>    |                                   |                                 |                |                |                |                |                                   |
| IM8801-2              | 0.512883 ± 14                     | 0.704436 ± 26                   | 8.52           | 39.2           | 12.7           | 755            |                                   |
| IM8801-5              | 0.512839 ± 14                     | 0.704370 ± 26                   | 9.13           | 43.0           | —              | 832            |                                   |
| IM8801-14A            | 0.512887 ± 14                     | 0.704447 ± 26                   | 8.97           | 41.2           | 13.0           | 841            |                                   |
| IM8808-5              | 0.512888 ± 16                     | 0.703941 ± 26                   | 6.87           | 29.7           | 40.7           | 755            |                                   |
| IM8808-10A            | 0.512890 ± 14                     | 0.703974 ± 26                   | 5.93           | 25.0           | 31.1           | 688            |                                   |
| IM8810-1A             | 0.512882 ± 14                     | 0.703667 ± 26                   | 11.3           | 54.9           | 47.2           | 1,043          |                                   |
| IM8814-1              | 0.512838 ± 14                     | 0.704308 ± 26                   | 9.81           | 47.1           | 28.8           | 847            |                                   |
| IM8814-2              | 0.512866 ± 14                     | 0.704529 ± 26                   | 8.65           | 41.5           | 33.2           | 790            |                                   |
| IM8815-7AB            | 0.512838 ± 14                     | 0.704288 ± 26                   | 6.18           | 27.6           | 29.6           | 730            |                                   |
| IM8815-7D             | 0.512839 ± 14                     | 0.704273 ± 26                   | 6.30           | 28.2           | 28.1           | 716            |                                   |

Isotopic data fractionation corrected using  $^{86}\text{Sr}/^{88}\text{Sr} = 0.1194$ ,  $^{148}\text{Nd}/^{144}\text{Nd} = 0.242436$ . NBS 987 Sr gives  $^{87}\text{Sr}/^{86}\text{Sr} = 0.710250 \pm 26$  and the La Jolla Nd standard gives  $^{143}\text{Nd}/^{144}\text{Nd} = 0.511859 \pm 14$ , where the uncertainties encompass the total range of measured standard values. Uncertainties given in the table are the greater of  $2\sigma$  standard error of the mean for the mass spectrometer run or the uncertainty for the standard. Equilibration temperatures were calculated using the method of Wells<sup>2</sup>.

Mongolia. We collected mantle xenoliths from basalts near the cities of Xilinhot and Abaga (Fig. 1) for isotope studies. Clinopyroxenes, and in some cases orthopyroxenes, were hand-picked from gently crushed and washed spinel peridotite xenoliths. Only interior portions, away from contact with the host basalt, were sampled. The hand-picked minerals were treated successively with purified water, 2 M HCl, 5% HF, 2 M HCl, and purified water, and finally hand-picked again to ensure high purity. Representative grains were retained for microprobe analyses and the remaining sample was processed for isotope analysis, as were host basalt powders. The Sr and Nd isotopic data are shown in Table 1. Procedural blanks analysed at the same time were well below levels that would influence the measured isotopic compositions, even for the low-concentration orthopyroxenes.

The isotopic compositions are typical for spinel peridotites from the continental lithosphere, with the exception of sample IM8815-7D, an otherwise unremarkable xenolith containing ~10% clinopyroxene with extremely high  $^{143}\text{Nd}/^{144}\text{Nd}$ . Virtually all samples have 'depleted' isotope ratios, as do the host rocks, which fall within the rather large range of the xenolith clinopyroxenes (Fig. 2). Sr and Nd isotopes show a well-defined negative correlation, and  $^{147}\text{Sm}/^{144}\text{Nd}$  and  $^{143}\text{Nd}/^{144}\text{Nd}$  are closely correlated on the isotope evolution diagram (Fig. 3).

Chemical data from electron microprobe analyses were used to determine equilibration temperatures for the xenoliths. Using the geothermometer of Wells<sup>2</sup>, we calculated temperatures ranging from 810 to 965 °C (Table 1). Although this geothermometer is probably accurate only to 50–70 °C, the relative temperature order may reflect a rough ordering in the depths at which these samples equilibrated.

Xenoliths with lower equilibration temperatures show a better correlation on the diagram of  $^{143}\text{Nd}/^{144}\text{Nd}$  against  $^{147}\text{Sm}/^{144}\text{Nd}$  than do those with higher temperatures, a feature which may

be related more to proximity in the lithosphere than to actual temperatures. If only those samples falling at or below the mid-point of the calculated range (that is,  $\leq 880$  °C) are regressed, they define an 'isochron' indicating an age of 1.64 Gyr. The initial ratio is  $0.510494 \pm 30$ , which is within error of the chondritic value. Even if all samples are regressed, the slope, and therefore the age, is the same within uncertainty, although the initial ratio is ~1 epsilon unit lower (see Fig. 3).

The evidence that the correlation in Fig. 3 has time significance is strong. Although there is a rough correspondence between isotope ratios and concentrations for both Sr and Nd, there is no linear relationship between  $^{143}\text{Nd}/^{144}\text{Nd}$  and  $1/\text{Nd}$ , a feature which argues against the isochron being a two-component mixing line. Furthermore, one potential mixing endmember, the host rocks, can be ruled out because the isochron passes well below this composition. A 'depleted' endmember would be required to have  $^{143}\text{Nd}/^{144}\text{Nd}$  greater than that of sample IM8815-7D, and thus would itself be ancient. The model age for sample IM8815-7D, relative to a chondritic mantle, is 1.64 Gyr, identical to the isochron age for the entire sample suite. Finally, as discussed above, restricting attention to samples with a narrow band of (low) equilibration temperatures, and thus presumably closely associated spatially within the mantle, substantially reduces scatter about the isochron. Although it is impossible to obtain field evidence concerning genetic associations for these mantle-derived samples, this feature implies a close relationship.

These arguments suggest that the observed correlation between  $^{143}\text{Nd}/^{144}\text{Nd}$  and  $^{147}\text{Sm}/^{144}\text{Nd}$  is a preserved lithospheric isochron, dating a thermal event during which there was melt generation and depletion of the lithosphere. The age given by the isochron would reflect the time at which this portion of the lithosphere cooled through an effective blocking temperature for this scale of sampling. This interpretation is only valid if

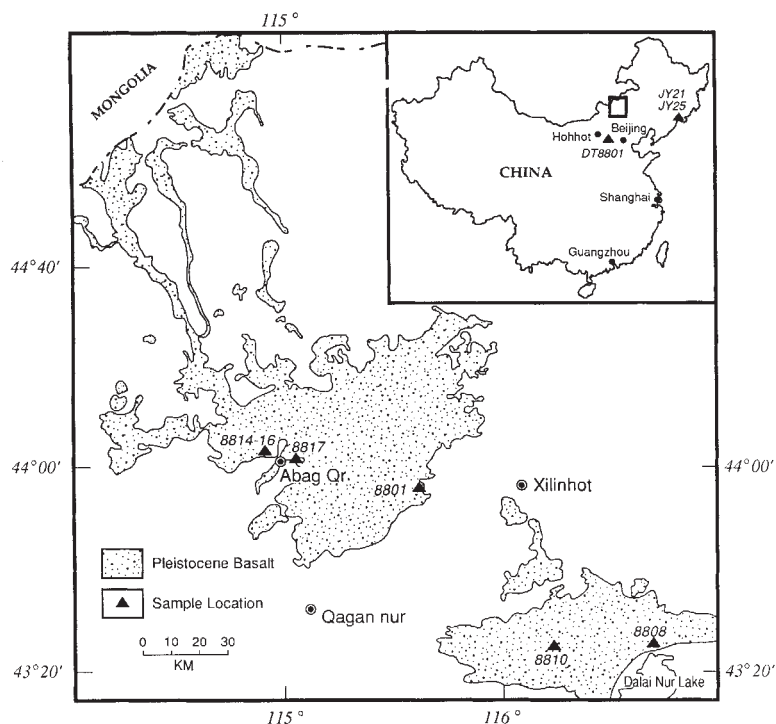


FIG. 1 Map showing the extent of Pleistocene basalt outcrops in the vicinity of Abaga (Abag Qr) and Xilinhot, Inner Mongolia. Sampling locations are marked with solid triangles.

the clinopyroxenes dominate the rare earth element budget, a reasonable assumption<sup>4</sup>. Diffusive processes act continuously to equilibrate isotopic compositions on a mineral scale, as is evident from data on two samples for which we have measured both clinopyroxene and orthopyroxene separates. In these xenoliths Nd isotopic compositions are equilibrated, or very nearly so, between the two phases (Table 1; Fig. 3).

If our interpretation is correct, discrete patches of the lithosphere under northeast China have retained their identity without significant chemical homogenization over 1.6 Gyr, while on a mineral scale equilibration has occurred. In the absence of metasomatic processes, this is plausible because, for simple volume diffusion, length scales for Nd diffusion should be much less than a metre over 1.6 Gyr. This assumes a diffusion coefficient around  $10^{-14} \text{ cm}^2 \text{ s}^{-1}$ , which is probably a

maximum for diopside at temperatures  $<900^\circ\text{C}$  (ref. 5). Our observations contrast with those of Song and Frey<sup>6</sup>, who reported chemical and isotopic features of peridotites from the Hannuoba region of Eastern China which indicate complex depletion and re-enrichment of the mantle at different times over at least the past 1 Gyr.

An initially puzzling aspect of the data is that the overall chemical compositions of the clinopyroxenes do not show a simple relationship with  $^{147}\text{Sm}/^{144}\text{Nd}$ , as might be expected if the latter is truly an index of the extent of depletion. Unfortunately, because the xenoliths are small, we do not have accurate information on their modal mineralogy that could be used to estimate bulk-rock chemical compositions. But microprobe data show that possible indicators of depletion, such as the molar ratio  $\text{Cr}/(\text{Cr} + \text{Al})$  in the clinopyroxenes, exhibit no relationship with  $^{147}\text{Sm}/^{144}\text{Nd}$ . Indeed, in sample IM8815-7D,  $\text{Cr}/(\text{Cr} + \text{Al}) = 0.0655$ , almost precisely in the middle of the range (0.0517 to 0.0798) for the suite as a whole. Ti concentration is lowest in IM8815-7D clinopyroxene, but in other samples shows no relationship with  $^{147}\text{Sm}/^{144}\text{Nd}$ . There is no hint in our samples of the correlation between  $^{143}\text{Nd}/^{144}\text{Nd}$  and basaltic components observed at Hannuoba and interpreted as evidence for re-enrichment of previously depleted mantle<sup>6</sup>. Only relatively small-degree partial melting, which would increase Sm/Nd ratios but have little effect on the overall major element composition of the clinopyroxene, seems consistent with the data presented here.

Because of the anomalous isotopic composition of sample IM8815-7D compared with all others analysed, we prepared a second separate for analysis. The results are similar to, but not identical with, those from the first analysis (Table 1, Fig. 3). Both Sr and Nd isotopic compositions differ between the two analyses by amounts considerably outside normal experimental uncertainty. Because of the low Nd concentration and extreme isotopic composition of this sample, we considered the possibility that blank Nd could account for the difference in  $^{143}\text{Nd}/^{144}\text{Nd}$ . But this would require more than 50 times our normal procedural blank contribution of  $\sim 1.5$  picograms (pg), and is extremely unlikely. The most probable explanation is that in spite of the careful leaching treatment, there is a small amount

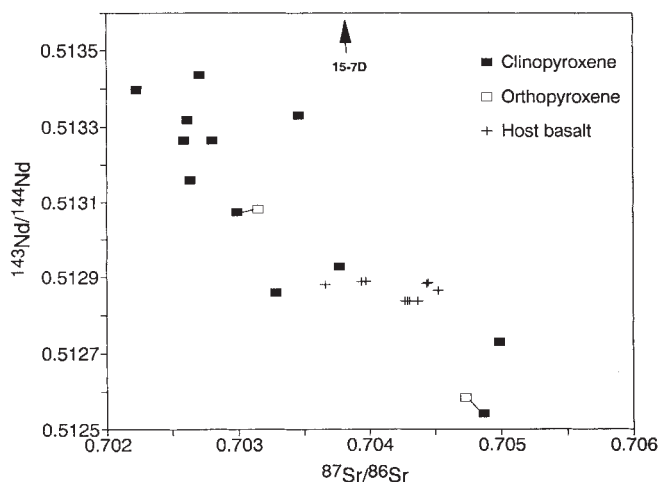


FIG. 2 Neodymium and strontium isotopic compositions of the clinopyroxenes are negatively correlated. Host rocks show a narrow spread in  $^{143}\text{Nd}/^{144}\text{Nd}$  and fall within the range of the clinopyroxenes. Coexisting clinopyroxene-orthopyroxene pairs are joined. Clinopyroxene from sample IM8815-7D falls at much higher  $^{143}\text{Nd}/^{144}\text{Nd}$  than all other samples.



of host-rock Nd incorporated into IM8815-7D-2. About 110 pg, the amount of Nd in  $\sim 4 \mu\text{g}$  of host rock, would be required to produce the observed shift. The form of the host-rock 'contaminant' is unknown, and it is possible that it affects both clinopyroxene separates from this sample. Because the effect is to move points roughly along the isochron, however, the slope of the regression would be little affected. The isotopic compositions of other clinopyroxenes analysed would not be changed measurably by the presence of comparable (or even much larger) amounts of a host-rock component because they are much closer in isotopic composition to their host rocks, and have Nd concentrations at least an order of magnitude higher than IM8815-7D (Table 1). The orthopyroxene separates, with low concentrations, could be affected and this may account for the slight differences in isotopic composition between orthopyroxene and clinopyroxene in the pairs we have measured (Table 1, Fig. 2).

There is much more scatter on the isochron diagram for Sr (not shown) than for Nd. In part this may result from analytical problems. Some of the clinopyroxenes have very low Rb concentrations (Table 1), and are therefore susceptible to the incorporation of even minute amounts of a host-rock component. Thus the higher concentrations in Table 1 should be viewed as upper limits. This is consistent with the fact that many of the separates with the lowest Rb contents (and therefore, presumably, minimum contamination) give Sr model ages between 1.4 and 1.8 Gyr, in accord with the Nd data.

The region in which these xenoliths occur is a Palaeozoic fold belt, sandwiched between two Archean continental nuclei: the Siberian platform to the north and the Sino-Korean platform to the south. Within the belt are fragments of ophiolite, indicating that an ocean basin or marginal sea once separated these blocks (see ref. 7 for a detailed discussion of the regional geology). It is thus perhaps surprising that the isotopic data do not reflect the mid-Permian age of this orogeny. Recent evidence, however, suggests that the main site of ocean closure during Permo-Carboniferous time may have been to the north of Inner Mongolia, and that the sediments now contained in the fold

belt may have been deposited on continental crust above an older suture zone<sup>7</sup>. The data reported here corroborate that suggestion, at least insofar as they present evidence for stable Proterozoic lithosphere beneath the fold belt. What then does the 1.6 Gyr age reflect? The Luliang orogeny<sup>8</sup>, a major episode affecting the Sino-Korean platform, probably occurred substantially earlier than the date recorded in the xenoliths. But abundant magmatic activity occurred in this region during the time span 1.5 to 1.7 Gyr (S.S. Sun, personal communication). Small but variable amounts of partial melting of mantle with roughly chondritic relative abundances of the rare earth elements during these events, leaving clinopyroxene in the source, would account for the present-day isotope characteristics of the xenoliths.  $\square$

Received 24 June; accepted 19 October 1992.

1. Menzies, M. A., Rogers, N., Tindle, A. & Hawkesworth, C. J. in *Mantle Metasomatism* (eds Menzies, M. A. & Hawkesworth, C. J.) 313–361 (Academic, London, 1987).
2. Wells, P. R. A. *Contrib. Miner. Petrol.* **62**, 129–139 (1977).
3. Brooks, C., Hart, S. R. & Wendt, I. *Rev. Geophys. Space Phys.* **10**, 551–577 (1972).
4. McDonough, W. F., Stosch, H.-G. & Ware, N. G. *Contrib. Min. Petrol.* **110**, 321–328 (1992).
5. Sneeinger, M., Hart, S. R. & Shimizu, N. *Geochim. cosmochim. Acta* **48**, 1589–1608 (1984).
6. Song, Y. & Frey, F. A. *Geochim. cosmochim. Acta* **53**, 97–113 (1989).
7. Mueller, J. F. et al. *J. Geol.* **99**, 251–263 (1991).
8. Yang, Z., Cheng, Y. & Wang, H. *The Geology of China*, Oxford Monogr. Geol. Geophys. **3**, 303 (Clarendon, Oxford, 1986).

ACKNOWLEDGEMENTS. We thank C. MacIsaac for assistance in the laboratory, and B.-Q. Zhu and X. Zhou for assistance with field work. Comments by S. S. Sun improved the manuscript. J.D.M. acknowledges the hospitality of Clare Hall and the Department of Earth Sciences, University of Cambridge. This work was supported by the U.S. NSF and the Academia Sinica, PRC.

## Foraging behaviour of emperor penguins as a resource detector in winter and summer

A. Ancel\*, G. L. Kooyman†, P. J. Ponganis†, J.-P. Gendner\*, J. Lignon\*, X. Mestre\*, N. Huin†, P. H. Thorson†, P. Robisson‡ & Y. Le Maho\*§

\* Centre d'Ecologie et Physiologie Energétiques, Centre National de la Recherche Scientifique, 23 rue Becquerel, 67087 Strasbourg, France

† Physiological Research Laboratory, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California 92093-0204, USA

‡ Centre d'Etudes Biologiques de Chizé, Centre National de la Recherche Scientifique, 79360 Beauvoir-sur-Niort, France

THE emperor penguin (*Aptenodytes forsteri*), which feeds only at sea, is restricted to the higher latitudes of the antarctic sea-ice habitat<sup>1–3</sup>. It breeds on the winter fast ice when temperatures are  $-30^\circ\text{C}$  and high winds are frequent<sup>3</sup>. Assuming entirely the task of incubating the single egg, the male fasts for about 120 days in the most severe conditions. When it is relieved by the female around hatching time, the distance between the colony and the open sea may be 100 km or more<sup>4,5</sup>, but where emperors go to forage at that time or during the summer is unknown. The polynias are areas of open water in sea-ice and during winter, with the under-ice habitats at any time of the year, they are among the most difficult of all Antarctic areas to sample. Here we monitor by satellite the routes taken by emperor penguins for foraging and compare them with satellite images of sea-ice. Winter birds walking over fast ice travelled up to 296 km to feed in polynias, whereas those swimming in light pack-ice travelled as far as 895 km from the breeding colony. One record of diving showed that although most dives are to mid-water depths, some are near the bottom. Obtaining such detailed information on foraging in emperor penguins means that this bird now offers a unique opportunity to investigate the Antarctic sea-ice habitat.

§ To whom correspondence should be addressed.

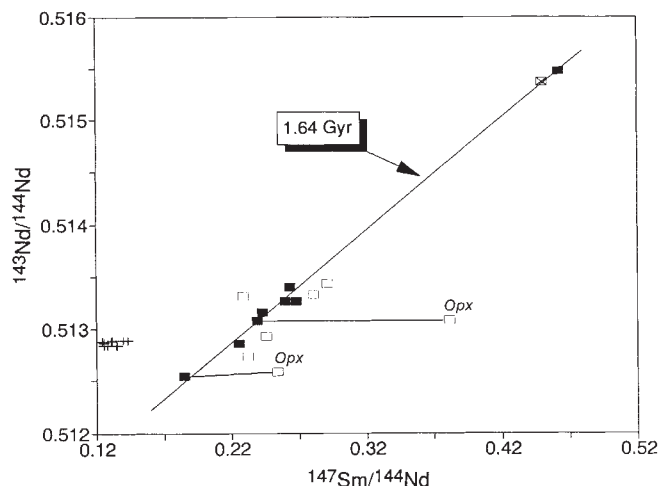


FIG. 3 On the isotope evolution diagram, clinopyroxenes are well correlated and define a 1.64 Gyr isochron (sample IM8815-7D-2, shown as a crossed open symbol, was not included in the regression, although its presence would not change the result). The initial ratio is the same as the chondritic (bulk Earth) value at that time. Solid symbols denote clinopyroxenes from xenoliths with equilibration temperatures  $\leq 881^\circ\text{C}$ . These show less scatter about the isochron than the samples with higher equilibration temperature. Two analysed orthopyroxenes are joined by solid lines to their co-existing clinopyroxenes. Host-rock compositions are shown as plus signs. The regression formalism of Wendt<sup>3</sup> was used in the isochron analysis. The slope and its formal uncertainty correspond to an age of  $1.642 \pm 0.014$  Gyr ( $2\sigma$ ) if all samples are included, and  $1.638 \pm 0.014$  Gyr if only those denoted by solid symbols are regressed.

TABLE 1 Tracking data

| Sex and bird number | Duration of tracking (days) | Number of reliable locations | Cumulated travelling distance (km) | Mean speed (km per h) | Outward journey |                  |                 | Return journey |                  |                 |
|---------------------|-----------------------------|------------------------------|------------------------------------|-----------------------|-----------------|------------------|-----------------|----------------|------------------|-----------------|
|                     |                             |                              |                                    |                       | Distance (km)   | Speed (km per h) | Number of halts | Distance (km)  | Speed (km per h) | Number of halts |
| M 1                 | 16                          | 43                           | 296                                | 0.8                   | 296             | 0.8              | 5               | —              | —                | —               |
| M 2                 | 7                           | 32                           | 82                                 | 0.5                   | >82             | —                | 4               | —              | —                | —               |
| M 3                 | 21                          | 67                           | 172                                | 0.3                   | >172            | —                | 5               | —              | —                | —               |
| M 4                 | 13                          | 55                           | 156                                | 0.5                   | 156             | 0.5              | 4               | —              | —                | —               |
| ? 5                 | 16                          | 76                           | 540                                | 1.4                   | 286             | 1.1              | 7               | 254            | 2.3              | 0               |
| F 12a               | 13                          | 65                           | 547                                | 1.7                   | 349             | 1.4              | 4               | 198            | 2.9              | 1               |
| F 12b               | 29                          | 107                          | 1,454                              | 2.1                   | 895             | 2.0              | 8               | 559            | 2.3              | 1               |
| M 13                | 16                          | 74                           | 760                                | 2.0                   | 496             | 1.5              | 3               | 264            | 3.0              | 1               |
| F 25                | 4                           | 21                           | 164                                | 1.7                   | 84              | 1.5              | 2               | 80             | 2.0              | 0               |
| ? 28                | 8                           | 30                           | 361                                | 1.9                   | 213             | 1.5              | 2               | 149            | 3.0              | 1               |

Data correspond to the tracking of four male emperor penguins overland during the winter (birds 1, 2, 3 and 4) and of five birds travelling by sea in the summer (birds 5 and 12 (two trips, 12a and 12b) 13, 25 and 28). F, female; M, male; ?, unknown.

Using the satellite Argos system (Fig. 1 legend), we tracked nine emperor penguins carrying transmitters for a total of 143 days and matched their locations to satellite images of sea-ice. Two groups of birds were studied under different circumstances. The first consisted of four males that were ending their winter fast. During the summer a second group consisting of both sexes were delivering food to a single chick upon their return. A time-depth recorder was also attached to one bird in this second group.

The first study was conducted in August 1990 at the Pointe Géologie colony, near Dumont d'Urville Station (66.7° S, 140.0° E) in Adélie Land. Based on satellite pictures at that time, clear open sea was at least 200 km away. During data collection, complete dark lasted for 11 out of 24 hours with no moonlight or bright auroras; twilight duration was 6 h. Although the

ambient temperature was about -30°C, which is below the design limits for the transmitter batteries, the birds were tracked for a distance ranging from 82 to 296 km (Table 1). They walked at any time over the 24 h at a mean speed of 0.5 km per hour. Two birds reached large polynias estimated at 110–130 km direct-path distance from the colony. The two other transmitters failed while the penguins were walking towards the eastern polynia (Fig. 1a). The cumulated distance the birds travelled on sea-ice was as much as 50% longer than the direct-path distance, and there were numerous satellite fixes (4 to 9 for each bird) at the same locations, indicating that these must be halt points. These halts only occurred between dusk and dawn and could last for that period, but they did not occur each night. As many as 55% of the halts were only 20 km away from the breeding colony. It was during these stop-overs that the birds occasionally

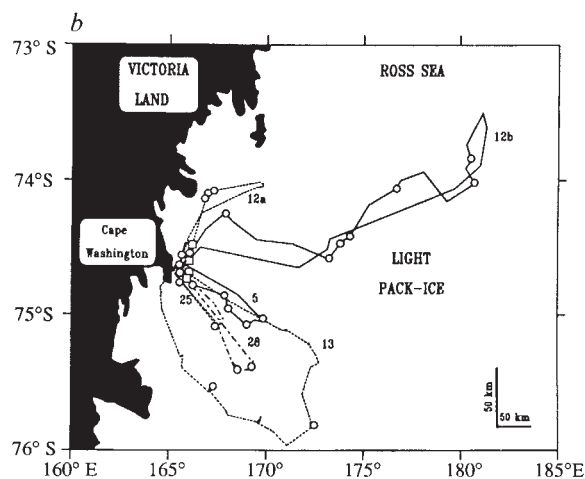
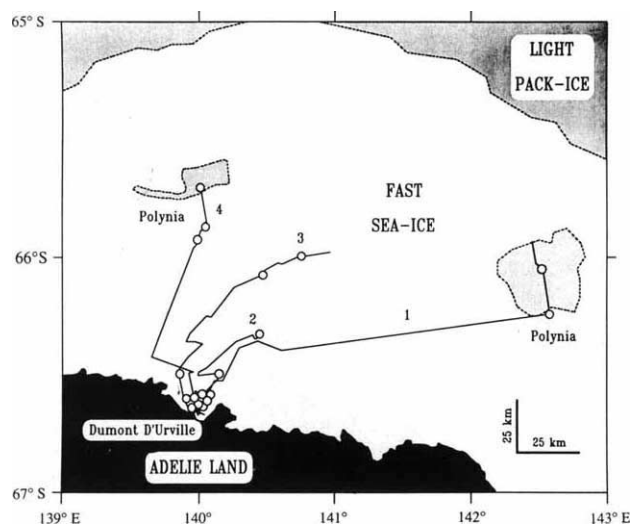


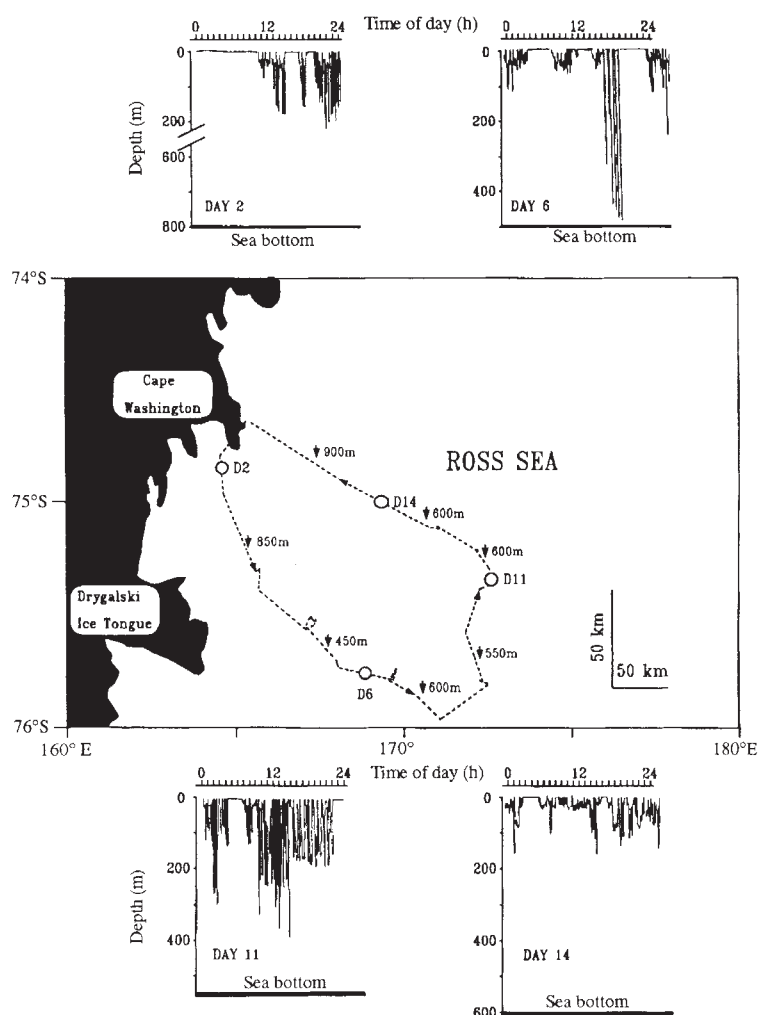
FIG. 1 a, Winter tracks of four (1–4) male emperor penguins on fast sea-ice matched with 11 satellite pictures of sea-ice. Direct path distances between the polynias are 130 km (north) and 110 km (east) from the Dumont d'Urville colony. Clear open sea is at least 200 km to the north of the colony. Broken dotted lines correspond to the mean extent of light pack-ice and polynias during the 3-weeks study period. Circles correspond to the stop-over points (halts) of the birds. b, Summer looping course of five emperor penguins (5, 12, 13, 25 and 28) travelling from and to Cape Washington, Ross Sea, during the summer chick-maturing period. Bird 12 went on two trips (12a and 12b). Circles and squares correspond to the halts of the birds during their outward and return journeys, respectively.

**METHODS.** The platform transmitter terminals (Toyocom T-2028 and Telonics ST-6) weighed between 245 and 475 g (includes transmitter, hydrodynamic housing, battery and antenna), which corresponds to 1–2% of the penguin's body mass. They were glued to the dorsal feathers of the birds using an

epoxy resin. Bird 13 also carried a 50-g Wildlife Computers time-depth recorder, which sampled dive depth every 5 s and had a depth resolution of about 4 m. The recorder stopped one day before the bird returned to the colony. On the return of the birds to the colony, transmitters were removed by peeling them from their feathers. The Argos system is based on two artificial satellites, NOAA, put into polar orbit at an altitude of 850 km, which provide 14 orbital rotations each per day, or between 22 and 28 data collections per day above the polar regions. When the passages of the satellites over the south pole regions coincide, this allows successive locations to be made in a very short time: on the same day, the movements of as many as six birds could be monitored from one point to another over only 2 to 5 min. Accuracy of location is to within less than 1 km. The resolution of satellite pictures of sea-ice was 1 km × 1 km. These were provided by P. Tildesley and H. Higgins (CSIRO, Hobart, Australia), and by R. Whritner (Scripps Institution of Oceanography).



FIG. 2 Satellite tracking of bird 13, with bathymetric data corresponding to location of dives. Compressed dive patterns are also shown for some days with the indication of depths for sea-bottom. Bathymetric map source: Antarctica, Ross Sea, Defense Mapping Agency Hydrographic Center, Washington DC (1978).



and over short periods travelled at higher speeds ( $3.8\text{--}10.6\text{ km h}^{-1}$ ) that were indicative of swimming, possibly to catch prey<sup>6</sup>. The emperor penguins must have had access to the sea through tide cracks, invisible on sea-ice pictures, around stranded icebergs kept open by glacier movement and through seal breathing holes<sup>7,8</sup>. Because of the many pauses, it took about 13–16 days to reach the polynias (Table 1).

The second study at the Cape Washington colony ( $74.6^{\circ}\text{S}$ ,  $165.4^{\circ}\text{E}$ ) in the Western Ross Sea lasted from late October to early December 1990. All birds were known to be successful breeders and were therefore shuttling between the sea and the colony to bring food to the chick. In contrast to the first investigation, there was open water of light pack-ice within 3 km of the colony. There were also 24 h of daylight. Satellite tracking showed that the birds covered between 164 and 1,454 km in a single foraging trip (Table 1). They generally followed a looping course (Fig. 1b), travelling faster during the return than during the outward journey (2.5 as opposed to  $1.6\text{ km h}^{-1}$ , respectively). The outward journey was interrupted by two to eight halts that lasted from 6 to 73 h. In contrast, the birds returned to the colony with only one major halt just before arriving at the colony. During the trip the birds had gained an average mass of 200 g per day.

From bird 13, a complete record of diving depths was obtained, part of which is shown in Fig. 2. When the dive bouts were overlaid on the track of the bird, it was clear that some dives were to the bottom or near it. On the outward journey the bird travelled about 70 km to the Drygalski Ice Tongue polynia, where open water occurs all the year round. Here it began feeding bouts by shallow diving to  $<150\text{ m}$  over a deep trench of 900 m. It then moved to a broad, shallower plain of about

500 m depth. Most dives were shallow ( $<50\text{ m}$ ), but occasional dives were over 400 m deep, to what must have been near the bottom. For example, on day 6 the penguin made a series of four dives ranging from 444 to 483 m over a bottom depth of about 450–500 m. For this bird and the other tracks, satellite images showed that the birds were diving and travelling in light pack-ice along the edge of a large open body of water to the south.

In conclusion, emperor penguins do not travel non-stop to reach foraging grounds that are far from their breeding colony. They are active during halts, indicating foraging. For birds walking on winter sea-ice, the halts only occur at night and in twilight. Birds moving in continuous daylight during summer may stop at any time. Finally, from satellite tracking, remote imaging and time-depth recording, it is apparent that the habitat used by breeding emperor penguins for feeding is extensive and, except for the very deepest areas, includes the entire water column.

Because the birds are specific feeders and bound to return to their chick and mate, they may be ideal bioindicators of some important southern ocean resources. As cheap oceanographic sampling platforms they may therefore provide information hitherto impossible to collect from oceanographic vessels because of costs, seasonal availability and sea-ice conditions. □

Received 11 May; accepted 28 September 1992.

1. Stonehouse, B. *Rep. Falkld Isl. Depend. Surv.* **6**, 1–33 (1953).
2. Le Maho, Y. *Am. Sci.* **65**, 680–693 (1977).
3. Prévost, J. *Ecologie du Manchot empereur* (Expéditions Polaires Françaises, Hermann, Paris, 1961).

4. Dewasmes, G. *et al.* *J. appl. Physiol.* **49**, 888–896 (1980).
5. Budd, G. M. *Proc. Zool. Soc., Lond.* **139**, 365–388 (1962).
6. Fraser, W. R. *et al.* *Polar Biol.* **10**, 37–41 (1989).
7. Kooyman, G. L. & Mullins, J. L. *Antarctic Ecosystems, Ecological Change and Conservation* (eds Kerry, K. R. & Hempel, G.) 169–176 (Springer, Berlin and Heidelberg, 1990).
8. Willing, R. L. *Nature* **182**, 194–195 (1958).

ACKNOWLEDGEMENTS. We thank E. Corbethau for technical assistance, J. Gilles for help with treating the data, D. Bresnahan for setting up air-support requirements, R. L. Gentry for the loan of Toyocom transmitters, and the 1990 winter team in Dumont d'Urville for field assistance. The research in Adélie Land was supported by grants from Terres Australes et Antarctiques Françaises, Centre National de la Recherche Scientifique and Université Louis Pasteur (Strasbourg). Logistic support was by Expéditions Polaires Françaises. The research at Cape Washington was supported by the NSF Division of Polar Programs.

## Functional localization and lateralization of human olfactory cortex

Robert J. Zatorre, Marilyn Jones-Gotman,  
Alan C. Evans & Ernst Meyer

McConnell Brain Imaging Centre, Montreal Neurological Institute and Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec H3A 2B4, Canada

**ANATOMICAL and physiological investigations in monkeys indicate that olfaction is subserved by several cortical regions<sup>1,2</sup>. But the areas implicated in the human olfactory system have not been definitively identified by functional criteria. Behavioural evidence<sup>3,4</sup> has suggested that laterally specialized mechanisms for odour processing may exist, but the neuroanatomical substrate remains unknown. We used positron emission tomography to study the cortical representation of human olfactory processing by comparing cerebral blood flow changes evoked during olfactory stimulation with those of a control task. We report here significant cerebral blood flow increases at the junction of the inferior frontal and temporal lobes bilaterally, corresponding to the piriform cortex, and unilaterally, in the right orbitofrontal cortex. The results complement and extend previous data implicating these regions in olfactory processing, and indicate that a functional asymmetry exists in the human brain favouring the right orbitofrontal area in olfaction.**

The neural basis for human olfaction is poorly understood. Indirect evidence from comparative studies suggests a structural and functional hierarchy, with piriform and orbitofrontal cortices corresponding to primary and secondary regions, respectively<sup>1,2</sup>. Experimental studies of human olfactory function after focal brain damage indicate that lesions of the temporal lobes produce consistent deficits in various olfactory tasks, while sparing odour-detection thresholds<sup>5–8</sup>, but the critical areas involved have not been precisely identified. Specific damage to either left or right orbitofrontal cortex impairs olfactory identification<sup>7</sup>, but deficits in odour discrimination are more pronounced after damage to the right orbitofrontal cortex<sup>3</sup>. The latter finding, together with the observation that odour discrimination performance is slightly but consistently better in the right than in the left nostril of normal subjects<sup>3,4</sup>, suggests a relative specialization favouring the right cerebral hemisphere for at least some olfactory tasks.

We sought to identify the cerebral structures activated by olfactory stimulation in normal human subjects in order to clarify the localization of primary and secondary olfactory cortices, and to investigate the possibility of lateralization of function. We measured cerebral blood flow (CBF) in 11 volunteers using the bolus H<sub>2</sub>O<sup>15</sup> technique<sup>9</sup> during two positron emission tomography (PET) scans. In the control task, subjects inhaled during birhinal presentation of an odourless cotton wand on each of eight trials. In the activation task, always done second, subjects smelled one of eight different odorants (Table 1) birhinally on each trial.

TABLE 1 List of odorants

| Odorant         | Descriptor       | Ratings     |              |           |
|-----------------|------------------|-------------|--------------|-----------|
|                 |                  | Familiarity | Pleasantness | Intensity |
| Shalimar        | Sweet perfume    | 3.0         | 3.7          | 3.4       |
| Butter extract  | Butterscotch     | 2.9         | 2.7          | 3.0       |
| Patchouli       | Woody, incense   | 3.1         | 3.0          | 3.8       |
| Kirsch          | Cherry           | 4.1         | 3.9          | 3.6       |
| Citronella      | Insect repellent | 4.5         | 3.8          | 3.5       |
| Hawes lemon oil | Kerosene         | 2.6         | 2.3          | 3.7       |
| Maple extract   | Maple syrup      | 4.6         | 4.1          | 3.4       |
| Lavender        | Flowery          | 3.5         | 3.2          | 3.4       |

Odorants used, giving verbal description of odour quality and mean ratings of pleasantness, familiarity and strength given by 30 normal subjects (none of whom participated as subjects in the PET study). Rating scales ranged from 1 to 5 (1=unpleasant, unfamiliar and weak intensity). Stimuli were presented undiluted on a cotton wand<sup>14</sup> and were chosen from a larger set on the basis of the ratings. Criteria for selection included that the items be moderately familiar but relatively difficult to name, distinct in quality and pleasantness from one another, and roughly matched for subjective intensity. Stimuli were presented sequentially at 7-s intervals throughout the 60-s scanning period.

TABLE 2 Positive activation foci observed during olfactory stimulation

| Focus | Coordinates (mm) |    |     | t-value | Area                             |
|-------|------------------|----|-----|---------|----------------------------------|
|       | x                | y  | z   |         |                                  |
| 1     | -21              | 6  | -21 | 5.52    | Left piriform cortex             |
| 2     | 20               | 6  | -15 | 4.13    | Right piriform cortex            |
| 3     | 17               | 39 | -13 | 3.66    | Right orbitofrontal cortex       |
| 4     | -7               | 5  | -11 | 3.50    | Left inferomedial frontal cortex |
| 5     | 21               | 6  | 15  | 2.73    | Right insula/claustrum           |
| 6     | -34              | -2 | 9   | 2.65    | Left insula/claustrum            |

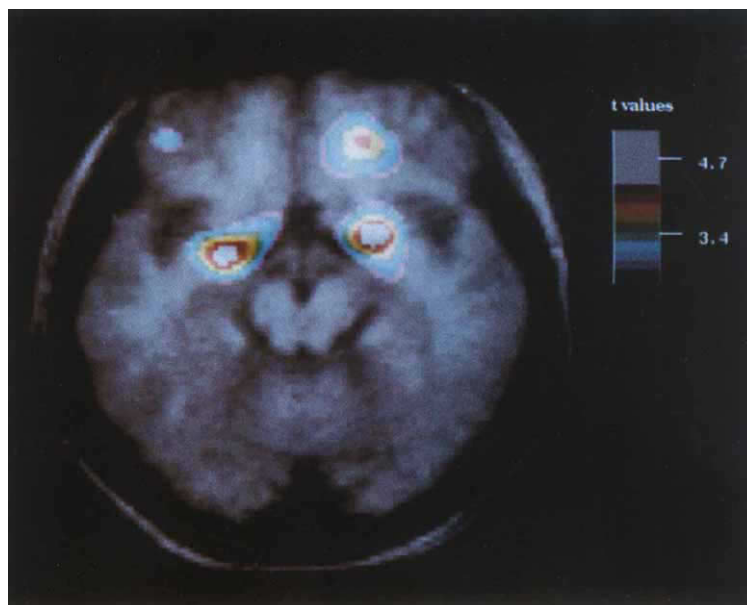
Stereotaxic coordinates<sup>15</sup> refer to medial-lateral position (x) relative to midline (positive=right), anterior-posterior position (y) relative to the anterior commissure (positive=anterior), and superior-inferior position (z) relative to the commissural line (positive=superior). The presence of significant focal changes was tested by a method based on three-dimensional gaussian random field theory<sup>16</sup>. Values equal to or exceeding a criterion of  $t=3.5$  were deemed statistically significant ( $P<0.0002$ , one-tailed, uncorrected). Correcting for multiple comparisons, a  $t$ -value of 3.5 yields a false positive rate of 0.58 in 200 resolution elements (each of which has dimensions  $20 \times 20 \times 7.6$  mm), which is roughly the volume of cortex scanned. For purposes of comparison, this  $t$  threshold corresponds to a Z-score of  $\sim 2.5$  ( $P=0.012$ , two-tailed) when using the change distribution analysis method<sup>17</sup>. The foci in the proximity of the right and left anterior insular cortex fall below the statistical criterion; however, it seems probable that they represent a true increase in CBF rather than an artefact, given their bilateral placement, and the known olfactory connections of the agranular and dysgranular portions of the insular region<sup>18</sup>. Another statistically nonsignificant ( $t=2.49$ ) area of activation was also noted in the left lateral orbitofrontal area (visible in Fig. 1), raising the possibility that some weak activation may also have been present on that side.

Stereotaxic coordinates for all significant positive changes are given in Table 2; the first three of these are depicted in Fig. 1. The two most significant foci are located at the junction of the temporal and inferior frontal lobes in each hemisphere, corresponding to the piriform cortex. The third focus is located in the right orbitofrontal cortex, corresponding roughly to Brodmann's cytoarchitectonic area 11. The fourth significant change (not visible in Fig. 1) was located in the left inferior medial frontal lobe.

The piriform cortex receives direct input from the ipsilateral olfactory bulb through the lateral olfactory tract in mammals, including humans<sup>2,10</sup>. Its strong activation in this study during olfactory stimulation confirms that this region corresponds

FIG. 1 Cortical regions activated by olfactory stimulation. The composite figure shows the averaged PET subtraction image superimposed on the averaged horizontal magnetic resonance image (MRI) scan, taken at a vertical level 17 mm below the commissural plane. Subtraction of the control from activated state yielded the focal changes in CBF shown here as a  $t$ -statistic image<sup>16</sup>. The range of  $t$ -values for the PET data is coded by the colour scale shown (see Table 2). The three areas of activation (corresponding to the first three foci listed in Table 2) are found within the piriform cortex bilaterally, and unilaterally in the right orbitofrontal cortex.

**METHODS.** Eleven right-handed subjects (five men, six women, mean age of 23 years) gave informed consent. PET scans were obtained with eyes open and ears plugged using the Scanditronix PC-2048 system<sup>19</sup>. The relative distribution of CBF was measured in baseline and activated conditions through the bolus  $H_2O^{15}$  methodology<sup>9</sup> without blood sampling<sup>20</sup>. Individual high-resolution MRI studies (63 slices, 2-mm-thick) were obtained from a Philips Gyroscan 1.5T and coregistered with the PET data<sup>21</sup>. An orthogonal coordinate frame was established based on the anterior-posterior commissure line as identified in the MRI volume<sup>22</sup>. These coordinates were used to apply a linear re-sampling of each matched pair of MRI and PET datasets into a standardized stereotaxic coordinate system<sup>15</sup>. PET images were reconstructed using a 20-mm Hanning filter, normalized for global CBF value, averaged across subjects, and the mean state-dependent CBF change image volume obtained<sup>23</sup>, which was then converted to a  $t$ -statistic volume<sup>16</sup>. Individual MRIs were subjected to the same averaging procedure, resulting in composite images for both  $t$ -statistic and MRI volumes. Anatomical and functional images were merged to allow direct localization on the MRIs of regions with a high  $t$ -value.



functionally to the primary olfactory cortex. The area of activity in the left medial frontal lobe (focus 4), adjacent to the left temporal piriform area (focus 1), may represent the portion of piriform cortex believed to lie within the frontal lobe, lateral to the olfactory tract<sup>10</sup>. The location and extent of the activation at the junction of the right frontal and temporal lobes (focus 2) suggest that it may subsume both frontal and temporal portions of the piriform cortex, obscuring any subdivision between the two.

The most striking finding was the unilateral activation of the right orbitofrontal cortex. This region receives important corticocortical projections from the piriform area<sup>1,2</sup>, as well as indirect olfactory input through the dorsal medial nucleus of the thalamus<sup>11</sup>. Recent theoretical attention has focused on the piriform cortex as a model system for recognition processes<sup>12</sup>, but our finding of orbitofrontal activation indicates that important neural processing must also take place in this region during higher-order olfactory processing. Consistent with this interpretation

is the fact that the physiological properties of neurons in orbitofrontal cortex are more selective than those of the piriform area<sup>1</sup>, and that orbitofrontal damage in humans is associated with major disturbances of odour identification and discrimination<sup>3,7</sup>. Future work must focus on the precise function of the several components of the cortical olfactory system identified here.

Despite the bilateral presentation of the stimuli, and the bilateral activation of piriform areas, the orbitofrontal cortex was significantly activated only on the right side. The lateralized blood flow increase suggests a functional specialization of the right orbital frontal cortex, in agreement with earlier behavioural findings<sup>3,4</sup>. Note also that the asymmetry arises at the level of the secondary, and not the primary olfactory cortex. This may be a general feature of the organization of laterally asymmetrical perceptual systems<sup>13</sup>, and suggests that hemispheric specialization is generally associated with higher-order processes, rather than initial sensory analysis. □

Received 7 May; accepted 12 October 1992.

1. Takagi, S. F. in *Cerebral Cortex* Vol. 9 (eds Peters, A. & Jones, E. G.) 133-152 (Plenum, New York, 1991).
2. Price, J. L. et al. in *Olfaction, a Model System for Computational Neuroscience* (eds Davis, J. L. & Eichenbaum, H.) 101-120 (MIT Press, Cambridge, Mass., 1991).
3. Zatorre, R. J. & Jones-Gotman, M. *Brain* **114**, 71-84 (1991).
4. Zatorre, R. J. & Jones-Gotman, M. *Percept. Psychophys.* **47**, 526-531 (1990).
5. Eskenazi, B., Cain, W. S., Novelly, R. A. & Friend, K. B. *Neuropsychologia* **21**, 365-374 (1983).
6. Eskenazi, B., Cain, W. S., Novelly, R. A. & Mattson, R. *Neuropsychologia* **24**, 553-562 (1986).
7. Jones-Gotman, M. & Zatorre, R. J. *Neuropsychologia* **26**, 387-400 (1988).
8. Eichenbaum, H., Morton, T. H., Potter, H. & Corkin, S. *Brain* **106**, 459-472 (1983).
9. Raichle, M. E., Martin, W. R. W., Herscovitch, P., Mintun, M. A. & Markham, J. *J. nucl. Med.* **24**, 790-798 (1983).
10. Eslinger, P. J., Damasio, A. R. & Van Hoesen, G. W. *Brain Cogn.* **1**, 259-285 (1982).
11. Potter, H. & Nauta, W. J. H. *Neuroscience* **4**, 361-367 (1979).
12. Haberly, L. B. & Bower, J. M. *Trends Neurosci.* **12**, 258-264 (1989).
13. Zatorre, R. J., Evans, A. C., Meyer, E. & Gjedde, A. *Science* **256**, 846-849 (1992).
14. Engen, T. *J. exp. Psychol.* **70**, 611-616 (1965).

15. Talairach, J. & Tournoux, P. *Co-Planar Stereotaxic Atlas of the Human Brain* (Thieme, New York, 1988).
16. Worsley, K. J., Evans, A. C., Marrett, S. & Neelin, P. *J. cereb. Blood Flow Metab.* **12**, 900-918 (1992).
17. Fox, P. T., Mintun, M. A., Reiman, E. M. & Raichle, M. E. *J. cereb. Blood Flow Metab.* **8**, 642-653 (1988).
18. Mesulam, M.-M. & Mufson, M. J. in *Cerebral Cortex* Vol. 4 (eds Peters, A. & Jones, E. G.) 179-226 (Plenum, New York 1985).
19. Evans, A. C., Thompson, C. J., Marrett, S., Meyer, E. & Mazza, M. *IEEE Trans. med. Imaging* **10**, 90-98 (1991).
20. Fox, P. T. & Raichle, M. E. *J. Neurophysiol.* **51**, 1109-1120 (1984).
21. Evans, A. C., Marrett, S., Torrescorzo, J., Ku, S. & Collins, L. *J. cereb. Blood Flow Metab.* **11**, A69-A78 (1991).
22. Evans, A. C. et al. *Neuroimage* **1**, 43-53 (1992).
23. Fox, P. T., Perlmutter, J. S. & Raichle, M. E. *J. Comp. Assist. Tomogr.* **9**, 141-153 (1985).

**ACKNOWLEDGEMENTS.** We thank the staff of the McConnell Brain Imaging Centre, its director, A. Gjedde, and the staff of the Medical Cyclotron unit of the Montreal Neurological Institute. Supported by grants from the Medical Research Council of Canada, the McDonnell-Pew Program in Cognitive Neuroscience, the Izaak Walton Killam Fund of the Montreal Neurological Institute, and the Fonds de la Recherche en Santé du Québec.



# Interactions between food-web structure and nutrients on pond organisms

Mathew A. Leibold\* & Henry M. Wilbur\*

Department of Zoology, Duke University, Durham, North Carolina 27706, USA

\* Present addresses: Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA (M.A.L.); Department of Biology and Mountain Lake Biological Station, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901, USA (H.M.W.)

THE suggestion that the densities of organisms are regulated by interactions involving entire food chains<sup>1</sup> has resurfaced into a new debate about the role of population regulation in environments that vary in productivity<sup>2-5</sup>. Past work has shown that density responses of organisms to enhanced productivity should be strongly affected by the number of trophic levels in the food chain<sup>6-11</sup> but more recent theoretical work on this question has suggested that heterogeneity within a trophic level can be just as important<sup>12-15</sup>. Here we present direct experimental evidence for such an interaction between environmental productivity and the species composition of herbivores. Our results illustrate that differences in food-web structure similar to those documented in natural communities<sup>16,17</sup> can dramatically alter how organisms will respond to abiotic factors such as nutrients that regulate primary productivity.

We manipulated food-web structure and nutrients in mesocosms that mimic natural ponds<sup>18,19</sup> to test for a food web-by-nutrient interaction on aquatic organisms. We created four food-webs by manipulating the occurrences of two important herbivores, *Rana utricularia* tadpoles<sup>20</sup> and *Daphnia laevis*<sup>15</sup> and manipulated their productivity by adding nutrients. We analysed data collected after the near-stabilization of population dynamics (except *Rana*) of the dominant organisms and used multivariate analyses of variance (Table 1) to test for interactions between nutrient levels and food-web structures.

Additions of nutrients significantly enhanced the biomass of zooplankton (Fig. 1), periphyton (Fig. 2a) and *Rana* ( $75.2 \pm 14.3$  g at low nutrients,  $152.2 \pm 33.8$  g at high nutrients, Student's  $t = 4.75$ ,  $P = 0.005$ ) demonstrating an enhancement of all trophic levels by the addition of nutrients. In addition, *Rana* suppressed periphyton (Fig. 2a) and *Daphnia* suppressed phytoplankton (Fig. 1), demonstrating significant negative impacts on their

respective resources. These direct effects on resources had additional indirect consequences; suppression of phytoplankton by *Daphnia* was associated with significant negative effects on other zooplankton (Fig. 1) and enhanced periphyton (Fig. 2) whereas suppression of periphyton by *Rana* was associated with a significant enhancement of phytoplankton densities (Fig. 1). Differences in phytoplankton density were strongly associated with differences in total water column nitrogen levels (Fig. 2b) suggesting that phytoplankton and periphyton compete for nutrients.

Most importantly, the long-term effects of nutrients on organisms depended qualitatively on manipulations of food-web structure as shown by significant interactions between food-web structure and nutrient level in the multivariate analysis of variance (MANOVA). Where *Rana* was present to control periphyton growth, nutrient additions enhanced phytoplankton biomass at least 10-fold (Fig. 1,  $P < 0.01$ , Tukey test comparing low and high nutrient treatments in the presence of *Rana*) but nutrient additions yielded reduced phytoplankton density when *Rana* were absent ( $P < 0.05$ , Tukey test comparing low and high nutrient treatments in the absence of *Rana*).

*Daphnia* also changed the responses of organisms to nutrient additions, but in a more subtle way than *Rana*. When *Daphnia* were present, phytoplankton blooms (caused by nutrient additions in the presence of *Rana*) consisted largely of *Scenedesmus*. But similar blooms in the absence of *Daphnia* consisted of micro-algae such as *Synechococcus* and unicellular *Microcystis* (Fig. 3). Thus *Daphnia* had significant negative main effects on the levels of phytoplankton, but it did not significantly alter the long-term effects of nutrients on phytoplankton. Instead it altered the specific composition of phytoplankton blooms.

Responses of zooplankton excluding *Daphnia* (primarily *Bosmina*, *Scapholeberis* and calanoid copepods) to nutrients also appeared affected by food-web structure (significant at  $P < 0.06$ , Table 1; Fig. 1) due to a complex interaction between the effects of *Rana* and *Daphnia*. Zooplankton density increased in response to nutrients significantly less when *Daphnia* was absent and *Rana* present than in any of the other three webs (Tukey test comparison of nutrient responses among the four food webs,  $P < 0.05$ ). No significant interactions between the structure of the food web and nutrient level were found for periphyton biomass although this may be due to measurement error associated with spatial heterogeneity.

Our results show that qualitatively different responses of organisms to enhanced productivity can depend on the structure of their food web as suggested by food-web models<sup>15</sup>. The effects of *Rana* were qualitatively important. In the presence of *Rana*,

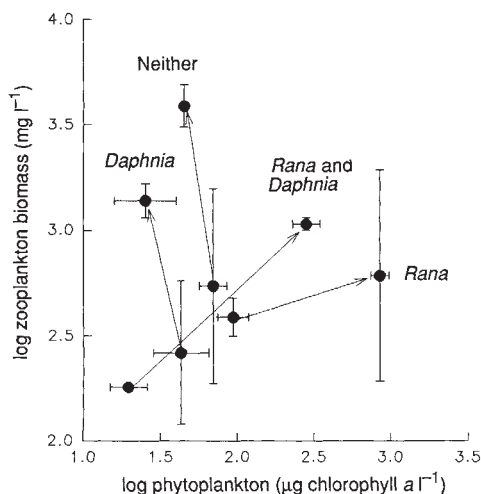


FIG. 1 Log<sub>10</sub> of phytoplankton density versus log<sub>10</sub> of zooplankton density (exclusive of *Daphnia*) in each food web at both nutrient levels. Arrows point from low nutrient to high nutrient conditions for each of four webs. Each arrow is labelled by the occurrence of *Rana* and *Daphnia*. Variation is indicated by bars representing standard deviations ( $n = 2-4$ ) for each response variable. Mesocosms (1,000 l) received 500 g washed pine straw as a benthic substrate, and were inoculated with a diverse assemblage of algae and zooplankton (excluding *Daphnia*). Hatchling *Rana utricularia* tadpoles (100) and 100 *Daphnia laevis* were introduced in all four combinations of occurrence. Populations were allowed to develop for 30 days until substantial zooplankton populations were apparent before the addition of nutrients to the high nutrient tanks (10 g  $\text{KH}_2\text{PO}_4$  and 65 g  $\text{NH}_4\text{NO}_3$ , in accordance with N:P ratios of the water; final N:P ratios varied slightly but were always less than 14 indicating nitrogen limitation). One tank where *Daphnia* were accidentally introduced before adding nutrients was recoded as a *Daphnia*-present replicate. Another tank, in which introduced *Daphnia* populations inexplicably did not establish, was deleted from subsequent analyses. Zooplankton samples, pooled from four standard locations in each mesocosm, were taken with a pipe-sampler and preserved in sucrose formalin. Biomass was calculated from density estimates and published length-weight regressions. Phytoplankton was measured by calibrated *in vivo* fluorescence from two pooled subsamples taken in each tank. Samples were averaged over a 20-day period to minimize variation due to oscillations in the density of some taxa in some replicates (our unpublished data).

TABLE 1 Statistical analysis of the effects of treatments on the biomass of phytoplankton, periphyton and zooplankton

| Effect             | (d.f.)      | Phytoplankton biomass ( $\mu\text{g l}^{-1}$ ) | Periphyton biomass ( $\mu\text{g cm}^{-2}$ ) | Zooplankton biomass ( $\text{mg l}^{-1}$ ) | MANOVA Roy's root |
|--------------------|-------------|------------------------------------------------|----------------------------------------------|--------------------------------------------|-------------------|
| Nutrient (N)       | (1) (3, 13) | 5.32 (0.0001)                                  | 1.82 (0.003)                                 | 11.98 (0.0002)                             | 5.47 (0.0001)     |
| Web                | (3) (3, 15) | 15.76 (0.0001)                                 | 6.52 (0.0001)                                | 10.00 (0.006)                              | 12.35 (0.0001)    |
| <i>Rana</i> (R)    | (1) (3, 13) | 8.24 (0.0001)                                  | 5.49 (0.0001)                                | 4.65 (0.009)                               | 8.53 (0.0001)     |
| <i>Daphnia</i> (D) | (1) (3, 13) | 4.81 (0.0001)                                  | 0.44 (0.09)                                  | 3.93 (0.02)                                | 3.51 (0.0002)     |
| D-by-R             | (1) (3, 13) | 0.89 (0.02)                                    | 0.01 NS                                      | 2.51 (0.05)                                | 0.71 (0.07)       |
| WEB-BY-N           | (3) (3, 15) | 11.83 (0.0001)                                 | 0.07 NS                                      | 4.77 (0.06)                                | 5.91 (0.0001)     |
| R-by-N             | (1) (3, 13) | 11.57 (0.0001)                                 | 0.01 NS                                      | 2.19 (0.06)                                | 5.80 (0.0001)     |
| D-by-N             | (1) (3, 13) | 0.04 NS                                        | 0.06 NS                                      | 1.39 (0.12)                                | 0.19 NS           |
| D-by-R-by-N        | (1) (3, 13) | 0.10 NS                                        | 0.01 NS                                      | 1.64 (0.10)                                | 0.29 NS           |
| Error              | (15)        | 2.06                                           | 2.11                                         | 7.88                                       |                   |

Entries for individual variables are sums of squared deviations attributed to each effect in an analysis of variation. Entry for MANOVA results are Roy's greatest root. Degrees of freedom for each effect are shown in parentheses; the first value is for the univariate tests, second pair is for the multivariate test. Two-tailed significance probabilities are in parentheses next to each entry (NS, probabilities greater than 0.2). Mean squared deviations and *F*-values of the test-statistic can be calculated from the figures given and are omitted for brevity. Results for food-web effects are presented as a total effect (labelled 'web') and partitioned into *Rana*, *Daphnia* and *Rana-by-Daphnia* subcomponent effects for both the main effects and the interaction effects with nutrients.

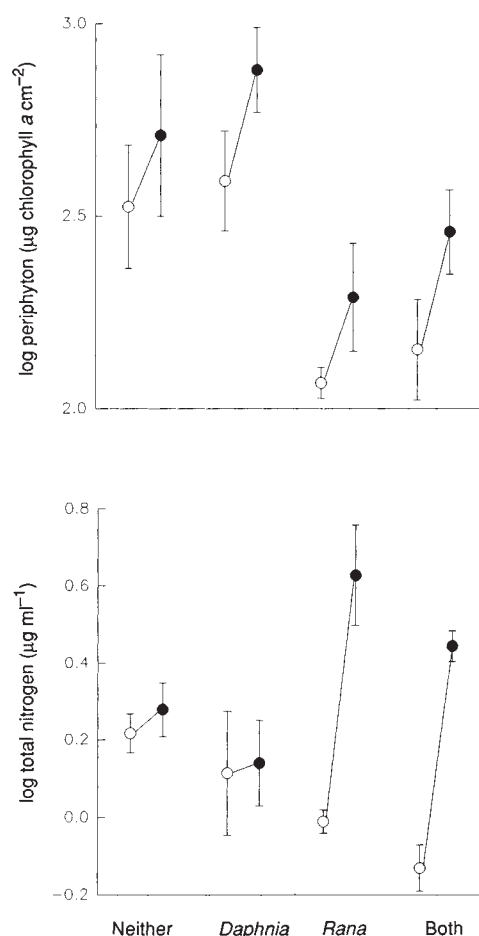


FIG. 2 Periphyton biomass as  $\mu\text{g cm}^{-2}$  chlorophyll *a* (Fig. 2a) and total water-column nitrogen (Fig. 2b) in each food web at both nutrient levels. Results for low nutrient levels are shown in open symbols, those for high nutrient levels are shown as closed symbols. Bars show  $\pm 1$  s.d. ( $n=2-4$ ) for each treatment. Periphyton was sampled twice at the end of the experiment from artificial substrates that had been incubated in the mesocosms for the duration of the experiment. Water for nutrient analyses was collected by slowly submerging a clean bottle through the water column of the mesocosm in an area free of periphyton. After persulphate digestion, total phosphorus was analysed by molybdate/ascorbic acid reduction and total nitrogen was analysed by hydrazine reduction methods on a TRAACS autoanalyser.

both phytoplankton and zooplankton increased in response to productivity, whereas zooplankton increased but phytoplankton decreased when *Rana* was absent. *Daphnia* had more subtle effects that altered the species composition of the phytoplankton blooms that result from nutrient additions in the presence of *Rana*. In both cases, the responses to nutrient additions of some species depended qualitatively on our manipulation of the structure of the food web.

Investigations of the effects of productivity on food webs have focused on the food-chain model as a heuristic tool<sup>3</sup> and have generally ignored how diversity within a trophic level might alter the responses of organisms to enhanced production<sup>21</sup>. Food-chain models predict that the population responses to productivity should depend on the number of trophic levels in the system. Our results demonstrate that effects due to the species composition of a specific trophic level (such as herbivores) can be just as substantial<sup>15</sup> and suggest that the entire array of interspecific interactions can potentially modify the population responses of organisms to environmental productivity. Such results illustrate that no single simple prediction of the effects

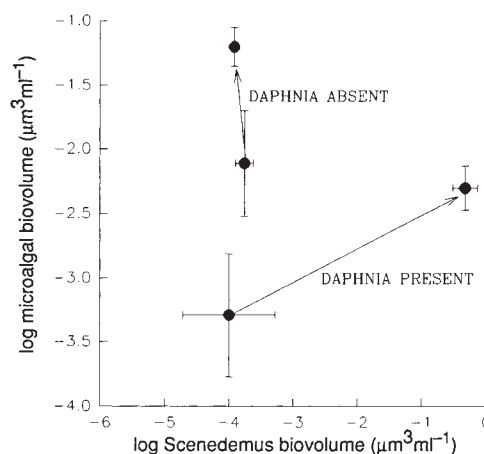


FIG. 3  $\log_{10}$  of microalgal biovolume versus  $\log_{10}$  *Scenedesmus* biovolume in the presence of *Rana*. Samples were collected as described in Fig. 1 and were preserved in acid Lugol's solution. Phytoplankton were enumerated and measured using the Utermohl method<sup>23</sup>. Food webs with and without *Daphnia* are indicated. Arrows point from low nutrient to high nutrient observations. Bars show  $\pm 1$  s.d. ( $n=2-4$ ) for the response variable for each treatment.

of nutrients on the abundances of organisms at different trophic levels can be made<sup>2-5</sup> and that future research might profitably focus on understanding mechanisms that produce such qualitative differences in the effects of productivity<sup>14,15,22</sup>. These results

also show that the topology of food webs as documented in recent studies<sup>16,17</sup> might have important consequences to the ways that the densities of component species are regulated by their abiotic environment. □

Received 6 July; accepted 24 September 1992.

1. Hairston, N. G., Smith, F. E. & Slobodkin, L. B. *Am. Nat.* **94**, 421-425 (1960).
2. Hunter, M. D. & Price, P. W. *Ecology* **73**, 724-732 (1992).
3. Power, M. E. *Ecology* **73**, 733-746 (1992).
4. Strong, D. R. *Ecology* **73**, 747-754 (1992).
5. Menge, B. A. *Ecology* **73**, 755-765 (1992).
6. Smith, F. E. in *Eutrophication: Causes, Consequences, Correctives* 631-645 (Natr. Acad. Sci. U.S.A., Washington DC, 1969).
7. Rosenzweig, M. L. *Science* **171**, 385-387 (1971).
8. Wolkind, D. V. *Am. Nat.* **110**, 431-447 (1976).
9. Oksanen, L., Fretwell, S., Arruda J. & Niemala, P. *Am. Nat.* **118**, 240-261 (1981).
10. Mittelbach, G. G., Osenberg, C. W. & Leibold, M. A. In *Size Structured Populations* (eds Ebenman B. & Persson, L.) 219-235 (Springer, New York, 1988).
11. DeAngelis, D. L. *Dynamics of Nutrient Cycling and Food Webs* (Chapman and Hall, London, 1991).
12. Phillips, O. M. *Arch. Hydrobiol.* **73**, 301-333 (1974).

13. Tilman, D. *Plant Strategies and the Dynamics and Structure of Plant Communities* (Princeton Univ. Press, Princeton, 1988).
14. Leibold, M. A. *Am. Nat.* **134**, 922-949 (1989).
15. Abrams, P. A. *Am. Nat.* (in the press).
16. Pimm, S. L., Lawton, J. H. & Cohen, J. E. *Nature* **350**, 669-674 (1991).
17. Cohen, J. E., Briand, F. & Newman, C. M. *Community Food Webs: Data and Theory* (Springer, New York, 1990).
18. Morin, P., Wilbur, H. M. & Harris, R. *Ecology* **64**, 1430-1436 (1983).
19. Wilbur, H. M. *Ecology* **68**, 1437-1452 (1987).
20. Seale, D. B. *Ecology* **61**, 1531-1550 (1980).
21. Oksanen, L. *Am. Nat.* **131**, 424-444 (1988).
22. Arditi, R., Perrin, N. & Saiah, H. *Oikos* **60**, 69-75 (1991).
23. Lund, J. W. G., Kipling, C. & LeCren, E. D. *Hydrobiology* **11**, 143-170 (1958).

ACKNOWLEDGEMENTS. We thank E. Walker for assistance with field and laboratory work and M. Power, L. Hedin and Gary Mittelbach for comments on the manuscript.

## Columns for visual features of objects in monkey inferotemporal cortex

Ichiro Fujita<sup>\*†</sup>, Keiji Tanaka<sup>\*‡</sup>, Minami Ito<sup>\*</sup> & Kang Cheng<sup>\*</sup>

<sup>\*</sup> Laboratory for Neural Information Processing, Frontier Research Program, RIKEN Institute, <sup>†</sup> Precursory Research for Embryonic Science and Technology (PRESTO), Research Development Corporation of Japan, and <sup>‡</sup> Information Science Laboratory, RIKEN Institute, Wako, Saitama 351-01, Japan

At early stages of the mammalian visual cortex, neurons with similar stimulus selectivities are vertically arrayed through the thickness of the cortical sheet and clustered in patches or bands across the surface. This organization, referred to as a 'column', has been found with respect to one-dimensional stimulus parameters such as orientation of stimulus contours<sup>1</sup>, eye dominance of visual inputs<sup>1</sup>, and direction of stimulus motion<sup>2</sup>. It is unclear, however, whether information with extremely high dimensions, such as visual shape, is organized in a similar columnar fashion or in a different manner in the brain. Here we report that the anterior inferotemporal area of the monkey cortex, the final station of the visual cortical stream crucial for object recognition<sup>3-8</sup>, consists of columns, each containing cells responsive to similar visual features of objects.

Most neurons in the anterior part of the inferotemporal cortex (IT) selectively respond to particular object features such as shape, or a particular combination of colour or texture with shape<sup>9-12</sup>. Anterior IT cells as a whole show a vast variety of preferred object features. If columnar organization exists, however, nearby cells should show a strong tendency to respond to similar features. We therefore first examined whether adjacent cells in the anterior IT respond to similar features or to distinct ones. We recorded extracellular action potentials from multiple cells with a single electrode, and activity of a single cell was isolated with a window discriminator. We determined the stimulus feature critical for activating this cell<sup>12</sup>, and devised a set of stimuli including the optimal and suboptimal stimuli, and their modifications, as well as ineffective stimuli, for this cell. This set was used to examine the responses of an adjacent cell or multiple cells selected with another window.

Adjacent cells at most sites responded together to the same stimuli in the set. At 41 of 47 recording sites (87%), simultaneously recorded cells showed maximal or nearly maximal responses to the same stimulus. At nine of the 41 sites, adjacent

cells behaved very similarly and no difference in stimulus preference was detected. At the other 32 sites, stimulus selectivities of adjacent cells were similar but differed in some aspects. At 21 of the 32 sites, neighbouring cells shared the most effective stimulus, but had a different selectivity to fine parameters. Two cells in Fig. 1, for example, responded to a vertical bar projecting from the top of a base, either a disk, a square or a horizontal bar (*a, b, e*). Components of the stimuli did not excite the cells (*c, d*). A small gap between the vertical bar and the base eliminated the responses (*f*). An 'L-shape' or a 'mirrored L-shape' evoked no or weak responses in the two cells (*g, h*). Both cells thus required, for maximal activation, the two adjoining corners at the junction of the projection and the base. Changing the intersection angles at the junction, however, revealed a difference between them. Cell 1 responded best to right angles (*e*) and maintained 50-60% of the maximal activation to other angles (*i, j*), whereas cell 2 responded only to right angles. In the other 11 of the 32 cases, the most effective stimuli differed slightly between adjacent cells, although they responded well to both stimuli: for example, one cell responded most strongly to a horizontally striped triangle, whereas the other preferred a square filled with horizontal stripes. Finally, at the remaining six sites, stimuli that activated one cell did not excite the other cell(s) simultaneously recorded. Neurons with similar selectivities thus tend to cluster together within the anterior IT.

We assessed the spatial extent of this clustering by inserting electrodes either vertically or obliquely to the cortex. As in the simultaneous recording experiments, we first determined the critical stimulus feature for a neuron, and made a set of stimuli including optimal, suboptimal and ineffective stimuli for that cell. We then tested other cells sampled at 100 or 200  $\mu\text{m}$  steps along the penetration with this same set.

In seven vertical penetrations where the angle between the electrode path and radial array of cells was 0-20°, the distance between the first recorded cell near the entry and the last recorded cell at the bottom ranged from 1.5 to 2.5 mm. Over a distance of 0.7-2.1 mm (mean  $\pm$  s.d.;  $1.3 \pm 0.5$  mm,  $n = 7$ ) within this thickness, we consecutively obtained cells that responded well to stimuli which were identical or similar to the optimal stimulus for the first cell tested in each penetration. The clustering covered 79-86% of the estimated cortical thickness (distance between first and last cell) in five of the seven penetrations, and 43-54% in the other two penetrations. The results indicate that cells with similar selectivities span most of the cortical layers.

In the example shown in Fig. 2, the first cell we examined (arrow) needed three stimulus features for its activation: (1) horizontally elongated overall shape; (2) upper part darker than the lower; and (3) darkest in centre part. We devised a set of



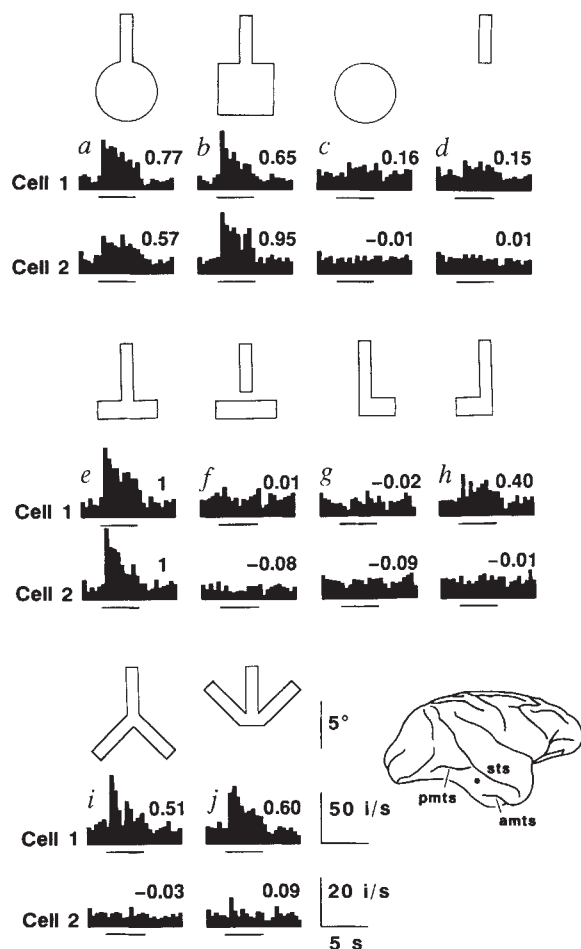
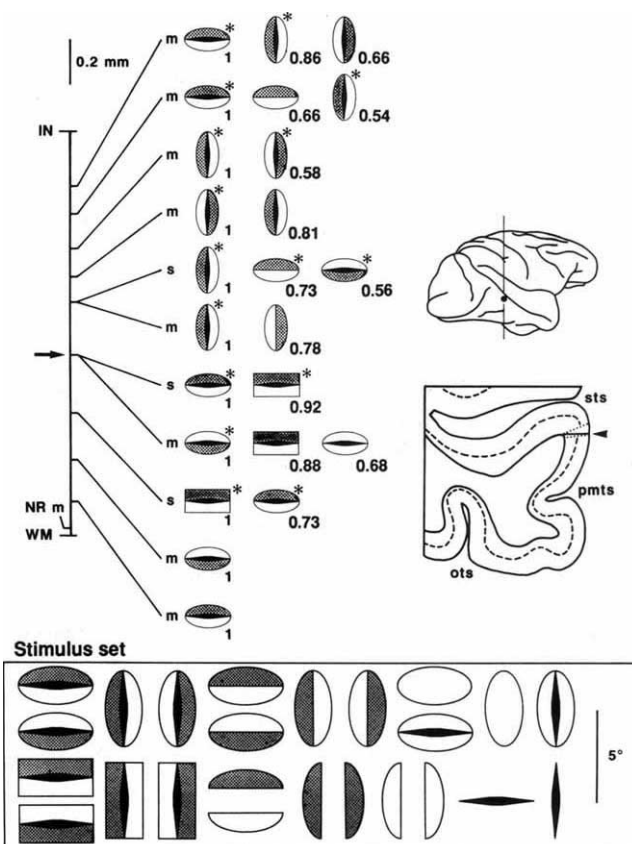


FIG. 2 Response profiles of 12 IT cells or multiple units recorded in a penetration directed vertically to the cortex. Location of the penetration is shown on the lateral view of the brain. The penetration is reconstructed on a coronal section through the IT (arrow head; ots, occipitotemporal sulcus). Dashed line indicates layer 4. Dotted lines indicate radial array of cells from which the penetration deviated by  $10^\circ$  at layer 4. A schematic drawing of the recording track (left) shows recording depths (IN, site where first neuron was recorded in this penetration; WM, entry into white matter). The first cell examined, seventh from the top (arrow), responded most strongly to the lips of a face among several tens of three-dimensional objects presented to the monkey. An extensive test revealed that the stimulus feature critical for activation was horizontally elongated shape with an upper half-dark, lower half-light pattern with the centre darkest. Neither colour nor contour shape of the lips was essential to activation. We then devised a set of 24 stimuli which included effective stimuli for this cell and their modifications (bottom), and tested other cells distributed throughout the grey matter in this penetration (s, single cell recording; m, multiple cell recording). Effective stimuli are shown to right of corresponding recording sites in order of the most to least effective stimuli, for statistically significant responses greater than 50% of the maximal response of each cell. None of the other stimuli in the stimulus set evoked significant responses in these cells. Responses were averaged over 10 stimulus presentations, and were considered significant when firing rate during stimulus presentation differed from the ongoing discharge rate just before presentation (Kolmogorov-Smirnov test;  $*P < 0.01$ ; no mark  $P < 0.05$ ). Numbers show response magnitude relative to the maximal response for each cell. All but one cell recorded over a distance of 1.3 mm responded only or maximally to stimuli with a half-dark, half-light pattern with the centre darkest. One multiple unit near the white matter did not respond to any of the stimuli in the set (NR).

**METHODS.** For vertical penetrations, electrodes were inserted into the brain from the side at right angles to midsagittal plane with or without a downward tilt ( $5$  or  $20^\circ$ ). In two of nine such penetrations we made, electrodes traversed in the bank of the superior temporal sulcus or of the anterior middle temporal sulcus, and the data were excluded from the analyses.

**FIG. 1** Two adjacent cells in anterior IT showing similar, but not identical, stimulus selectivity. The recording site is shown on the lateral view of the cortex. amts, pmts, Anterior and posterior middle temporal sulcus; sts, superior temporal sulcus. Stimuli were white solid shapes on black background. Impulse histograms below each visual stimulus were averaged over 10 stimulus presentations. Horizontal lines indicate 4-s periods of stimulus presentation. Numbers show response magnitude relative to the maximal response. Negative values indicate suppression below the ongoing firing rate. Both cells responded to two adjoining corners at the junction of a base and a vertical projection, but had different selectivity to intersection angles.

**METHODS.** General experimental procedures were similar to those in ref. 12. Recordings were made in area  $ITd^{23}$  of six monkeys (*Macaca fuscata*) prepared for repeated recordings<sup>12</sup>. In recording experiments, a small hole was made in the skull for electrode insertion under initial anaesthesia with ketamine followed by isoflurane. During the recording, the monkey was immobilized, and anaesthesia was maintained with a mixture of  $N_2O$  and  $O_2$ . Electrocardiogram helped us to judge anaesthetic depth, and sodium pentobarbital was supplemented when necessary. Extracellular activity was recorded simultaneously from multiple cells with single electrodes. Responses of one cell isolated from multiunit activity were tested with a set of hand-held stimuli (see ref. 12) for effective stimuli survey. An image of the most effective object was taken with a video camera and stored on a computer. Stimuli were moved with a small amplitude (up to  $1.2^\circ$ ) around the centre of the predetermined receptive field on a TV monitor. With a computer graphics system, we simplified the effective stimuli step by step to determine stimulus features essential to cell activation. We first changed the image from full colour to monochrome. If the monochromatic image was as effective as the original one, we did not pursue the cell's selectivity to colour. If the original image was better, we examined colour selectivity of the cell. We next removed texture, gradual luminosity change, and highlight from the image one by one to determine whether any of these features was essential. If not, we modified the image to a regular geometrical figure; for example contours of an apple were changed into a disk with a projecting bar. If the cell's response survived this transformation, we then determined the critical aspect of the figure by modifying shapes of components or overall configuration or by decomposing the image into components. Throughout these procedures, impulse histograms were generated for the isolated cell and for a nearby cell or multiple unit, which were selected by a second window discriminator. Electrolytic lesions were made at some recording sites to reconstruct electrode tracks, and were later verified in Nissl-stained sections.



24 stimuli, and tested other cells along this penetration. All but one cell recorded over a distance of 1.3 mm (80% of estimated thickness of grey matter) responded exclusively or maximally to stimuli with a half-dark, half-light pattern with the centre darkest. The darkest centre or a white ellipse alone did not evoke responses in any of the tested cells.

In another vertical penetration (Fig. 3), cells recorded over a distance of 2.1 mm (86% of estimated thickness of grey matter) responded only to stimuli with a gradual change of luminosity. Thus the preferred stimulus feature shared by vertical arrays of cells was not restricted to shapes, but could be other visual features of objects such as gradation.

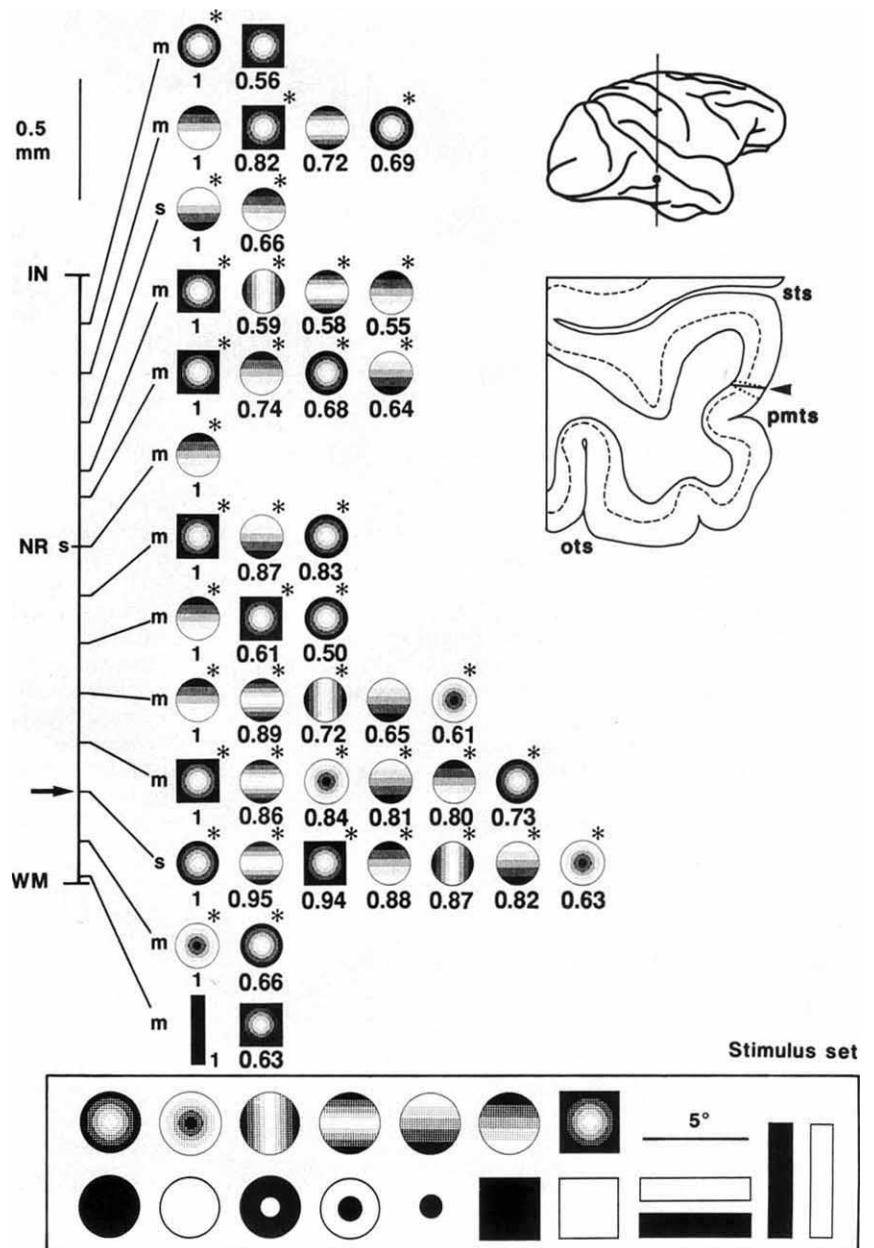
Along tangential or oblique penetrations where the angle formed by the recording track and radial array of cells was 50–90°, cells with similar stimulus selectivity were localized within 0.2–0.7 mm ( $0.4 \pm 0.2$  mm,  $n = 7$ ), a shorter distance than that obtained in vertical penetrations ( $P < 0.001$ , Student's  $t$ -test). This indicates that the clustering was vertically elongated in the direction of cortical depth and was patchy across the surface. Neurons immediately outside the cluster showed no

response to stimuli that evoked maximal or substantial responses in the first tested cell. In six of the seven tangential penetrations, we found another cell or cluster of cells showing similar stimulus preferences after a gap of 0.4–1.0 mm. In three of the six recording tracks, there was a third cluster after a similar gap. Cells outside the clusters were visually responsive, and their preferred stimulus features were likely to be distinct from those activating the clusters, because some of them responded to stimuli in the test set which did not activate cells in the clusters or to some objects presented to the monkey.

Figure 4 shows an example in which a cluster of four single or multiple units localized within 0.2 mm and another cell 0.46 mm away responded to a combination of two different colours. Shape was not a critical factor, and we used rectangles of two colours in the test set. A component colour alone evoked no or weaker responses. Neither a combination of black and white nor equiluminant grey activated the cells. A cell outside the cluster responded to a star shape which did not activate the cells in the cluster.

The similarity of cells along a penetration was due to local

FIG. 3 Response profiles of 14 anterior IT cells recorded in a vertical penetration to the cortical surface. The angle between the penetration and radial arrangement of cells was 20° at layer 4. The stimulus feature essential for activation was first determined for the third cell from the bottom in the figure (arrow). Other neurons in this penetration were tested with a set of 18 stimuli (bottom) which included the optimal, suboptimal and ineffective stimuli for the first examined cell. Although actual stimulus figures presented to the monkey had 256 grey levels, they are shown only with 5 grey levels for clarity. All cells except two responded only to stimuli with a gradual change of luminosity. Two exceptions were one cell near the white matter which preferred a black vertical bar over a square with a gradation, and the other recorded in the middle of the penetration (indicated as NR), which showed responses to a gradation stimulus (upper row, fifth from left, in the set) in 5 of 10 presentations, but the response did not reach a significant level.

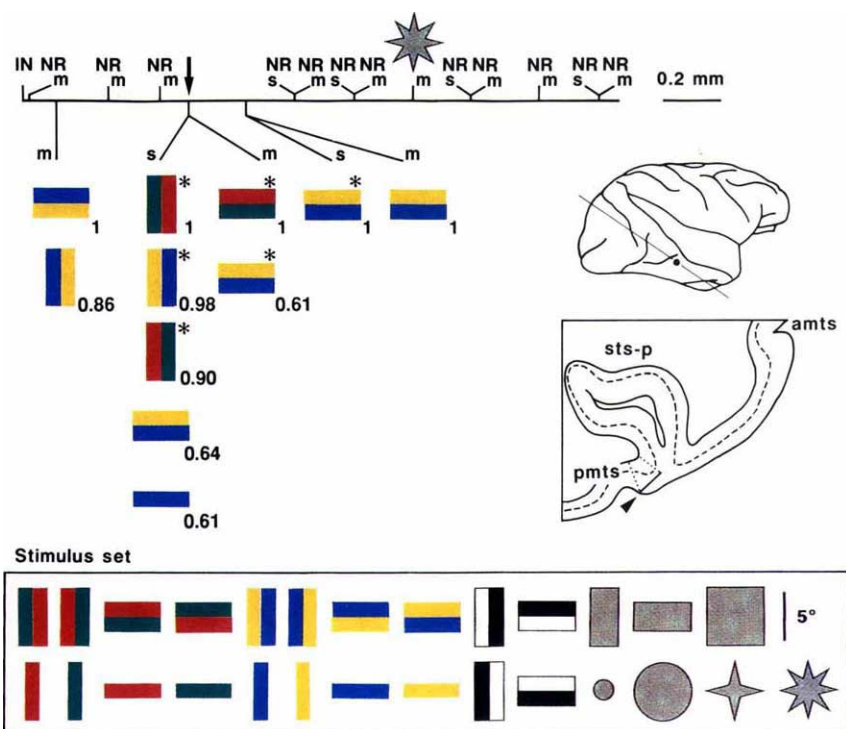




**METHODS.** For oblique or tangential penetrations, electrodes were directed anteriorly into the brain from the side at an angle of  $45^\circ$  to midsagittal plane and downwards by  $27\text{--}30^\circ$  (see inset).

We conclude that anterior IT neurons selective for similar object-features are aligned normal to the cortical surface across most, if not all, cortical layers. Previous studies have shown clustering of cells with similar response properties in the IT<sup>12,13</sup>, a cortex in the superior temporal sulcus<sup>14,15</sup>, and the medial superior temporal cortex<sup>16</sup>. But evidence for columnar organization is fragmentary, or analysis of neuronal stimulus selectivity is insufficient in these studies. Our results provide, to our knowledge, the first evidence for the existence of functional columns in a higher association cortex, and together with results in lower sensory cortices of various modalities<sup>1,2,17,18</sup>, support the notion that columnar organization is a basic feature of the cerebral neocortex<sup>19–21</sup>.

Although neurons in an IT column responded to similar



The surface area of the anterior IT dorsal to the anterior medial temporal sulcus (TEd<sup>23</sup>), from which our recordings were made, is  $330 \pm 76 \text{ mm}^2$  ( $n = 4$ ; our unpublished observation). Cells with similar selectivity were localized in clusters of  $0.4 \text{ mm}$  width on average. If we take a square of this width ( $0.16 \text{ mm}^2$ ) as the area of an individual column, this gives us 2,000 as a rough estimate of the number of columns in this area. The number of distinct object features represented in this area would be smaller than this number because we observed multiple clusters of cells with similar selectivities along tangential penetrations. Studies on this limited number of distinct object features common to vertical arrays of cells may help us to understand what the 'basis functions' are in the neural coding of object images. □

1. Hubel, D. H. & Wiesel, T. N. *J. Physiol.* **160**, 106-154 (1962).
2. Albright, T. D., Desimone, R. & Gross, C. G. *J. Neurophysiol.* **51**, 16-31 (1984).
3. Zeki, S. & Shipp, S. *Nature* **335**, 311-317 (1988).
4. Felleman, D. J. & Van Essen, D. C. *Cerebral Cortex* **1**, 1-47 (1991).
5. Mishkin, M., Ungerleider, L. G. & Macko, K. A. *Trends Neurosci.* **6**, 414-417 (1983).
6. Gross, C. G. in *Handbook of Sensory Physiology* (ed. Jung, R.) 451-482 (Springer, Berlin, 1973).
7. Dean, P. in *Analysis of Visual Behavior* (eds Ingle, D. J., Goodale, M. A. & Mansfield, J. R. W.) 587-628 (MIT Press, Cambridge, 1982).
8. Iwai, E. *Vision Res.* **25**, 425-439 (1985).
9. Gross, C. G., Rocha-Miranda, C. E. & Bender, D. B. *J. Neurophysiol.* **35**, 96-111 (1972).
10. Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. J. *J. Neurosci.* **4**, 2051-2062 (1984).
11. Baylis, G. C., Rolls, E. T. & Leonard, C. M. *J. Neurosci.* **7**, 330-342 (1987).
12. Tanaka, K., Saito, H., Fukuda, Y. & Moriya, M. *J. Neurophysiol.* **66**, 170-189 (1991).
13. Gochin, P. M., Miller, E. K., Gross, C. G. & Gerstein, G. L. *Expl Brain Res.* **84**, 505-516 (1991).
14. Perrett, D. I. *et al. Hum. Neurobiol.* **3**, 197-208 (1984).
15. Perrett, D. I. *et al. Behav. Brain Res.* **16**, 153-170 (1985).

16. Saito, H. *et al. J. Neurosci.* **6**, 145-157 (1986).
17. Abeles, M. & Goldstein, M. H. *J. Neurophysiol.* **33**, 172-187 (1970).
18. Mountcastle, V. B. *J. Neurophysiol.* **20**, 408-434 (1957).
19. Mountcastle, V. B. in *The Mindful Brain* (eds Edelman, G. M. & Mountcastle, V. B.) 7-50 (MIT Press, Cambridge, MA, 1978).
20. Kaas, J. H. in *Signal and Sense, Local and Global Order in Perceptual Maps* (eds Edelman, G. M., Gall, W. E. & Cowan, W. M.) 67-82 (Wiley-Liss, New York, 1990).
21. Van Essen, D. C., Felleman, D. J., DeYoe, E. A., Olavarria, J. & Knerim, J. *Cold Spring Harb. Symp. quant. Biol.* **55**, 679-696 (1990).
22. Marr, D. & Nishihara, H. K. *Proc. R. Soc. B* **E200**, 269-294 (1978).
23. Iwai, E. & Yukiie, M. *Brain Res.* **444**, 397-401 (1988).

**ACKNOWLEDGEMENTS.** We thank C. G. Gross and M. P. Stryker for comments on the manuscript and T. Fujita for technical assistance. This study was supported by Frontier Research Program, Science and Technology Agency, Japan; Precursory Research for Embryonic Science and Technology, Research Development Corporation of Japan; and grants from the Japanese Ministry of Science, Culture and Education.



# *Drosophila wingless* generates cell type diversity among *engrailed* expressing cells

Scott Dougan & Stephen DiNardo

The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

**DURING embryogenesis, body pattern is established in a stepwise process<sup>1</sup>. After specification of the body axis, the embryo is subdivided into smaller units. Within these units, a diverse array of cell types is then generated. The subdivisions of the *Drosophila* embryo, called parasegments<sup>2</sup>, are defined by the interface between cells expressing the homeoprotein Engrailed and cells expressing the secreted protein Wingless<sup>3-5</sup>. We have examined the generation of cell-type diversity within parasegments by focusing on the choice of cell fate made by the *engrailed* (*en*)-expressing cells. These cells differentiate as one of two alternative cell types<sup>6</sup>. We report here that this choice is mediated by *wingless* (*wg*), in a function distinct from its early role maintaining *en* expression<sup>7,8</sup>. Thus, *en* cells exhibit different responses to the *wg* signal at different developmental stages. Early *wg* input stabilizes the subdivision of the body axis by maintaining *en* expression, whereas later input generates cell-type diversity.**

On differentiation, each epidermal cell secretes a cuticular structure, the shape of which reflects its fate<sup>9</sup> (Fig. 1a). Ventrally, cells either secrete smooth cuticle or hair-like structures, called

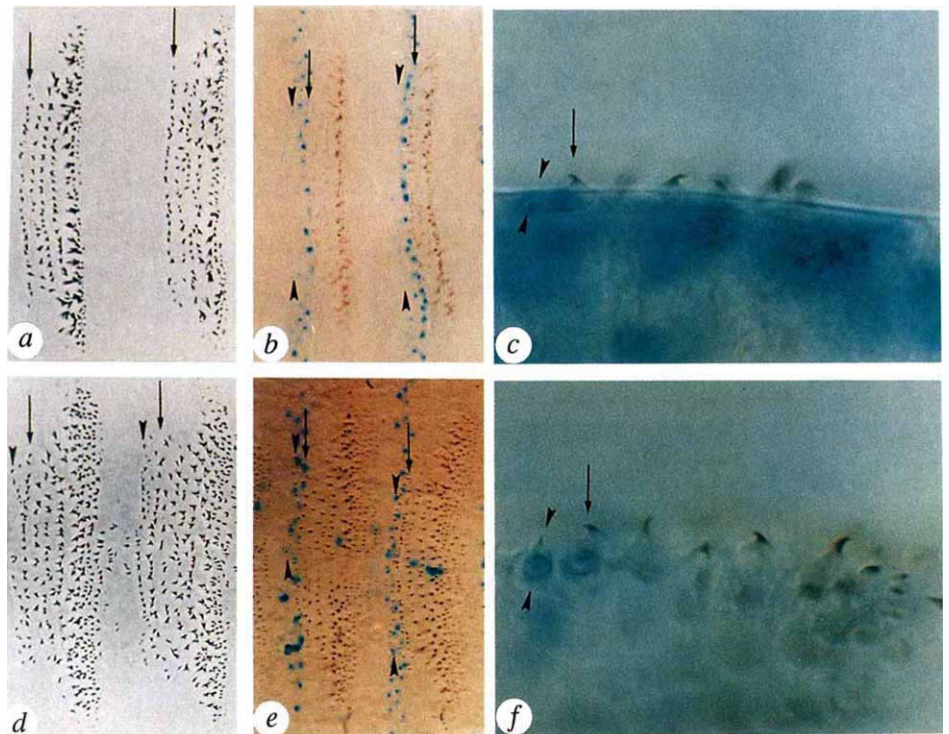
denticles. Denticles are arranged in rows, with about six rows comprising a denticle belt. Each row of denticles is of characteristic size and orientation<sup>9</sup>. To visualize *en* cell fate, we used a *lacZ* reporter gene controlled by the *en* promoter<sup>10</sup>. *en* is expressed in stripes roughly two to three cells wide (Fig. 1b). Anterior *en* cells underlie smooth cuticle (Fig. 1b, c, arrowheads) whereas the posterior *en* cells underlie first row denticles, which are small and hook anteriorly (Fig. 1b, c, arrows). Thus, cells expressing *en* can adopt either the smooth cell fate or a first row denticle cell fate<sup>6</sup>.

We tested whether *wg* function, which is required for the production of smooth cuticle<sup>11,12</sup>, directs this choice. Inactivation of a thermosensitive *wg* allele after 8 h of development results in the production of an extra row of small denticles anterior to the normal first row (Fig. 1d)<sup>11</sup>. This suggests that the anterior *en* cells have incorrectly selected a denticle fate. Expression of the *en-lacZ* reporter gene directly shows that both the normal first row denticles and the extra row arise from *en* cells (Fig. 1e, f, arrows and arrowheads). Thus, *wg* input is required after 8 h to instruct the anterior row of *en* cells to choose the smooth cell fate. Inactivating *wg* after 9.5 h no longer affects the choice of *en* cell identity<sup>11</sup> (data not shown). Therefore, the fate of the anterior *en* cells is specified by *wg* between 8 and 9.5 h.

Unlike the anterior *en* cells, the fate of the posterior *en* cells was not affected by elimination of late *wg* function. Because posterior *en* cells adopt the denticle fate independently of *wg*, they might not receive the signal or, alternatively, cannot respond to it. The ability of the posterior *en* cells to respond to the *wg* signal can be tested by expressing *wg* on both sides of

FIG. 1 The fate of anterior *en* cells is dependent on late *wg* function. Cuticle patterns from the fourth and fifth abdominal segments of wild type (a-c) and *wg<sup>ts</sup>* (8-h inactivation) (d-f). b, e, ventral views, d, f, lateral views of the cuticle in relation to *en* expression. Arrows indicate the normal position of the first row denticles. Anterior is to the left in all figures. a, Wild type. b, *en*-expressing cells visualized by  $\beta$ -galactosidase activity produced by an *en-lacZ* reporter construct that encodes an unstable, nuclear *en-lacZ* fusion protein<sup>7,10</sup>. *en* is expressed in a 2-3 cell wide stripe, with anterior cells (arrowheads) more weakly stained than posterior ones (arrows). c, Optical section of the cell types comprising a denticle belt, stained as in b. Denticles representing rows 2, 5 and 6 are in a different focal plane. Among the slightly flattened cells underneath the cuticle, the anterior (arrowheads) and the posterior (arrow) *en* cells are visible. The *en*-expressing cells in CNS are out of focus, causing the other blue stain. d, *wg<sup>ts</sup>*; an extra row of small denticles (arrowheads) is produced anterior to the normal first row (arrow). Other cells also produce small denticles<sup>11,12</sup> which are apparent between the denticle belts. e, *en* cells in a *wg<sup>ts</sup>* embryo; visualized as in b. Because *en* expression becomes independent of *wg* at 6 h<sup>11,16</sup>, it is properly maintained in these embryos. f, *wg<sup>ts</sup>*, optical section of a denticle belt; stained as in b. Both the anterior (arrowheads) and the posterior (arrows) *en* cells produce denticles.

**METHODS.** Cuticles were prepared<sup>14</sup> and photographed by phase contrast. The *ts* allele *wg<sup>ts</sup>* was used in all experiments<sup>19</sup>. Gastrulae were selected under oil and allowed to develop at a permissive temperature (18 °C) until they reached the developmental equivalent of 8 h at 25 °C. Embryos were checked to verify that they had reached the appropriate stage (mid-stage 12), and shifted to 25 °C for the remainder of embryogenesis. The *en-lacZ*

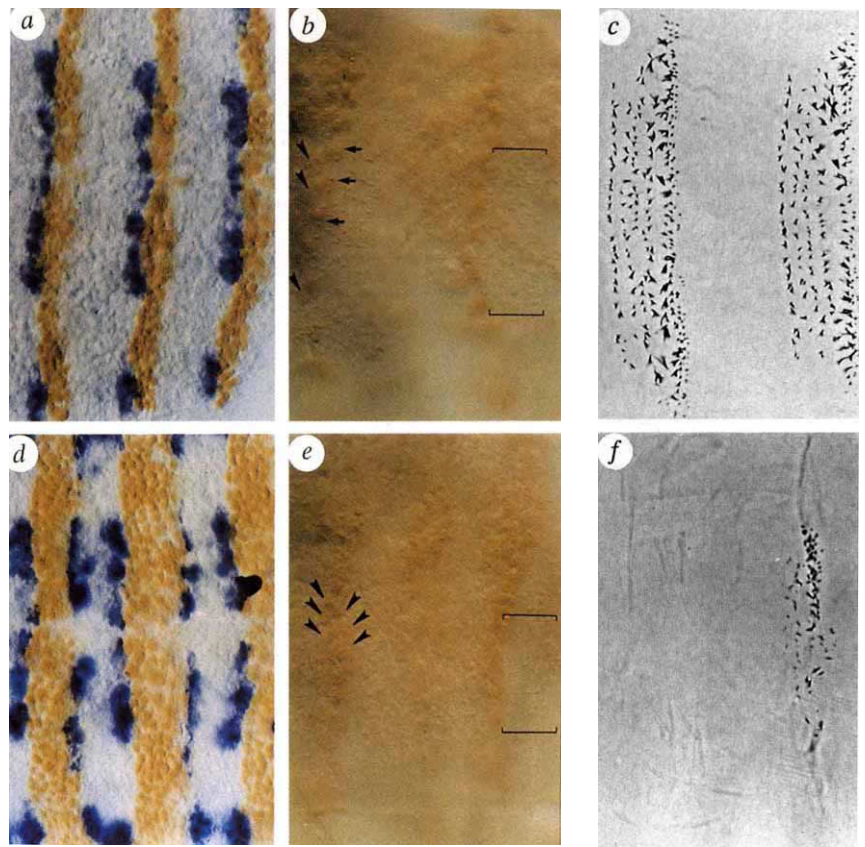


construct (F4)<sup>10</sup> was recombined onto the *wg<sup>ts</sup>* chromosome. Embryos (22-24 h old) were dechorionated, hand devitelinated, and fixed 30 min in a 1:1 mix of heptane and 25% glutaraldehyde in 50mM sodium cacodylate. Fixed larvae were decapitated with a tungsten needle to allow efficient penetration of substrate. The  $\beta$ -galactosidase reactions<sup>20</sup> were incubated for two days at room temperature. Embryos were post-fixed in 4% formaldehyde in PBS with 0.1% Tween-20, and put through a series of 30%, 50% and 80% glycerol in PBS. For b and e, the CNS and gut were dissected away. Intact embryos, c and f, were mounted laterally.



FIG. 2 Posterior *en* cells adopt the smooth fate in *nkd* mutants. Ventral *en* expression in wild-type (a, b) or *nkd* mutant (d, e) abdominal parasegments is shown in relation to *wg* expression at 5.5 h (a, d) and alone, at the onset of cell differentiation at 13 h (b, e). a, Wild type; expression of *wg* RNA (blue) and Engrailed protein (brown). b, *en* expression (brown nuclei) in wild type. The anterior *en* cells (arrowheads) are not associated with shape changes. The posterior *en* cells (arrows) form small protrusions. Larger protrusions are apparent from the cells posterior to the *en* cells (brackets), identifying cells fated to form the five other denticle rows. c, Wild-type cuticle. d, *nkd* mutant; stained as in a. *wg* is induced posterior to the broadened *en* stripe at about 5 h. e, *en* expression in *nkd*. Protrusions are not observed from either the anterior or posterior *en* cells (arrowheads). No protrusions are evident between *en* stripes, indicating that these cells either die before differentiation or become smooth cells. Occasionally some denticle cell types differentiate (brackets). These serve as an internal control, indicating that differentiation occurs at the proper time in *nkd* mutants, but that cell fates are mis-specified. The width of *en* stripes varies at this stage, and occasionally *en* expression fades from a stripe (see middle stripe), possibly indicative of cell death. But crossing an *en-lacZ* fusion construct into *nkd* mutants reveals that many *en* cells do survive well beyond differentiation (data not shown). Detailed analysis of these stainings was prevented by the diffuse  $\beta$ -galactosidase activity generated within the abnormal *nkd* gut. f, *nkd* cuticle. Most cells adopt smooth cell fates, although occasionally remnants of denticle belts are observed.

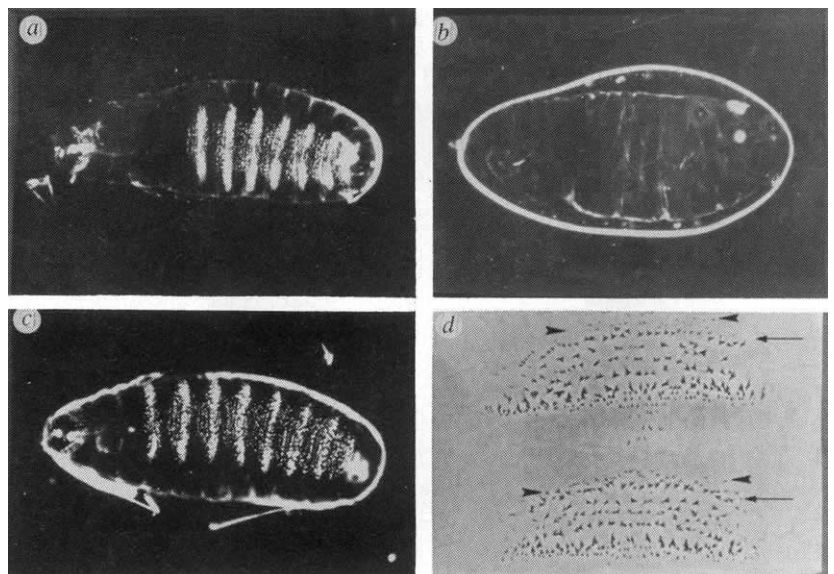
METHODS. *nkd*<sup>TE</sup> was used in all experiments<sup>14</sup>. After the *in situ*<sup>21</sup> hybridization, using an RNA probe, embryos were dehydrated through 30%, 50% and 70% ethanol, 5 min each rinse, and then left overnight at 4 °C before rehydration. The embryos were then processed for immunocytochemistry<sup>7</sup>



and filleted in 80% glycerol. Stage 16 embryos<sup>22</sup> processed for immunocytochemistry against Engrailed were dehydrated, dissected in methyl salicylate/canada balsam to remove the gut and CNS, and mounted ventral side up. Photographs with Nomarski optics.

FIG. 3 Inactivation of late *wg* function restores denticle cell types to *nkd* embryos. Dark-field images of ventral cuticles (a-c). a, *nkd*; *wg*<sup>ts</sup> (shift-up at 6 h). Embryos resemble identically treated *wg*<sup>ts</sup> single mutants (see c), indicating that the rescue is complete. b, *nkd* (shift-up at 6 h). *nkd* single mutants are unaffected by the temperature shift demonstrating that the rescue of denticles observed in the double mutants is due to the late inactivation of *wg*. c, *wg*<sup>ts</sup> (shift-up at 6 h). d, *nkd*; *wg*<sup>ts</sup> (shift up at 8 h), sixth and seventh abdominal segments. Anterior is to the top. Arrows indicate the normal first row denticles. Later inactivation of *wg* in *nkd* mutants results in the duplication of first row denticles (arrowheads), showing that in *nkd*, as in wild type<sup>11</sup> (see Fig. 1d), the fate of anterior *en* cells is dependent on late *wg* input. A time course of *wg* inactivation in *nkd* mutants revealed that rescue of denticle belts occurs between 6 and 9.5 h (data not shown), showing that as in wild type<sup>11</sup>, the role of *wg* specifying cell fates in *nkd* mutants has ended by 9.5 h.

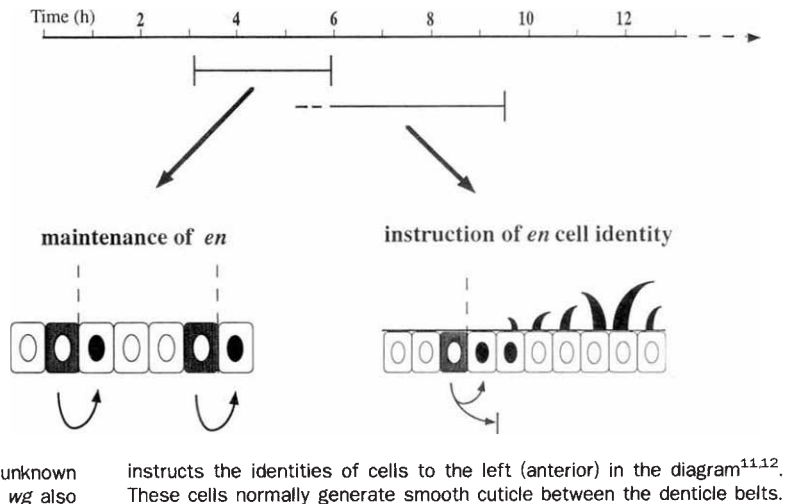
METHODS. To identify the *wg*<sup>ts</sup>; *nkd* double mutants, we used a *nkd* chromosome marked with *Ubx*, which transforms the first abdominal segment to the third thoracic segment. *wg*<sup>ts</sup>; *nkd* double mutant embryos were generated as follows: Siblings from a cross of *wg*<sup>ts</sup>/CyO to *nkd Ubx*/TM1 *Me* flies were collected. *wg*<sup>ts</sup>/+; *nkd Ubx*/+ brothers and sisters were crossed. Embryos were selected and treated as described in Fig. 1. Of 332 embryos, 211 were wild type; 49 *nkd*; 49 *wg*<sup>ts</sup>; 23 *wg*<sup>ts</sup>; *nkd Ubx*. This ratio is close to the predicted ratio 9:3:3:1. Controls in which the embryos were raised at the restrictive temperature yielded 85 wild type; 27 *nkd*; 35 *wg*<sup>ts</sup>.



This is close to the prediction if the *wg*<sup>ts</sup>; *nkd* double mutants fall into the *wg*<sup>ts</sup> class. Of 388 embryos raised at the permissive temperature, 108 (28%) died, none with a *wg*<sup>ts</sup> phenotype. This indicates that *wg*<sup>ts</sup> homozygous embryos hatch when raised at 18 °C. The *wg*<sup>ts</sup> phenotype at various times of *wg* inactivation and the recombination frequency between the *nkd* and *Ubx* loci (7%) were determined by crossing *wg*<sup>ts</sup>/+; +/TM1 *Me* parents, and of crossing *nkd Ubx*/+; +/CyO parents, respectively.



FIG. 4 Two distinct responses of *en* cells to *wg* input. Cells form at 2.5 h and epidermal cells differentiate at 13 h. From 3–6 h (solid line), *wg* is required for the maintenance of *en* expression<sup>7,8,11,16</sup>. A schematic parasegment at the cellular blastoderm stage (3 h). The dashed lines represent the parasegmental boundaries. A signal (curved arrows), probably the wingless protein, is produced by the *wg* cells (grey cytoplasm), allowing the maintenance of *en* expression (filled nuclei), which is expressed at this stage in a single-cell-wide stripe. Between 6 and 9.5 h, the *wg* signal instructs the identities of the anterior *en* cells. A schematic parasegment about 10 cells wide at late stages, is drawn. The final identities, first visible at the onset of cell differentiation (13 h), are diagrammed above the cells. At this stage *en* is expressed in two-cell-wide stripes. *wg* instructs the anterior *en* cells to adopt the smooth fate (curved arrow below the cells). This decision is made between 8 and 9.5 h. We propose that the cells which normally give rise to the six denticle fates are prevented from receiving the *wg* signal during normal development (indicated by curved arrow blocked by a line). This permits a second, unknown process to generate the denticle cell types. Between 6 and 9.5 h, *wg* also



the *en* stripe. Such mis-expression of *wg* is observed in *naked* (*nkd*) mutants<sup>8</sup>. In wild type, *wg* RNA is expressed in the row of cells anterior to the two cell wide stripe of *en* cells (Fig. 2a)<sup>8,13</sup>. In *nkd* mutants, extra cells express *en* forming an abnormally broad stripe (Fig. 2d). Ectopic *wg* RNA expression is then induced in the row of cells just posterior to the broad *en* stripe (Fig. 2d)<sup>8</sup>. On differentiation, exclusively smooth cell types are generated (Fig. 2f)<sup>14</sup>. Because many cells die in *nkd* mutants, we directly determined *en* cell fates by staining with antibodies against Engrailed at the onset of cell differentiation (13 h), when cells form molds of the denticles to be secreted<sup>15</sup>. In wild type, the anterior *en* cells exhibit no shape changes (Fig. 2b, arrowheads), indicative of their smooth fate. The posterior *en* cells, however, form small protrusions reflecting their denticle fate (Fig. 2b, arrows). By contrast, none of the *en* cells in *nkd* mutants manifest shape changes, indicating that they adopt the smooth cell fate (Fig. 2e, arrowheads).

We tested whether this change of fate is due to late function of *wg*. If the late ectopic expression of *wg* in *nkd* mutants causes posterior *en* cells to adopt the smooth cell fate, then eliminating *wg* activity should allow the posterior *en* cells to adopt their normal denticle fate. But if the fate change is due to another defect in *nkd* embryos, such as the broadened domain of *en*, then removing *wg* activity from *nkd* mutants should have no effect. *wg* function was eliminated from *nkd*, *wg*<sup>ts</sup> double mutant embryos by shifting to the non-permissive temperature at 6 h of development. This treatment restores denticles to the double mutants (Fig. 3a). The broadened domains of *en* remain indistinguishable from *nkd* single mutants through 10 h (data not shown), suggesting that the rescue is not a secondary effect of the loss of *wg* on the regulation of *en*. Thus, in *nkd* mutants, the posterior *en* cells respond to late *wg* input, adopting the smooth cell fate. This raises the possibility that these cells also have the capacity to respond to *wg* in wild type, but are not normally exposed to it.

In *nkd* embryos, the smooth fate is also adopted by cells that do not express *en*, but ordinarily give rise to denticle rows 2–6. In wild type, these cells form protrusions at the onset of differentiation (13 h) corresponding to the future denticle rows 2–6 (Fig. 2b, bracket). In *nkd* mutants, cells posterior to the *en* cells do not exhibit shape changes. Either these cells adopt the smooth fate in *nkd*, or they do not survive to differentiate (Fig. 2e). Without molecular probes for these cells, we cannot distinguish between these possibilities. But by removing late *wg* function, all denticle cell types are restored to *nkd* mutants (Fig. 3d). This suggests that the cells normally giving rise to denticle rows 2–6 respond to the late *wg* signal in *nkd* mutants, and raises the possibility that these cells, like the posterior *en* cells, are normally prevented from receiving this signal.

instructs the identities of cells to the left (anterior) in the diagram<sup>11,12</sup>. These cells normally generate smooth cuticle between the denticle belts.

In summary, *en* cells in the ventral epidermis adopt one of two alternative fates and late *wg* input mediates this choice. This and previous work<sup>7,8,11,16</sup> demonstrates that there are two distinct periods during embryogenesis when *en* cells respond to the *wg* signal (Fig. 4). Early *wg* input maintains *en* expression, stabilizing the parasegmental boundaries, whereas later *wg* input instructs the fate of the anterior *en* cells, generating cell type diversity within the parasegments. A function analogous to the early requirement for *wg* has been found for the mouse homologue Wnt-1<sup>17</sup>. It is required for the maintenance of cells expressing the mouse engrailed homologue, En-1. Like *wg*, Wnt-1 may also have a second, later function in the generation of cell type diversity among the En-1 cells.

The mechanism by which the six denticle cell types are generated is unknown. Because these identities cannot be generated in the presence of extra *wg*, as occurs in *nkd* mutants, a central feature of this mechanism must be the restriction of the late *wg* signal itself, or of a cell's ability to respond to this signal. Consistent with this, studies of protein localization show that at late stages Wingless is distributed in an asymmetric gradient within every parasegment, forming a sharp posterior boundary<sup>4,18</sup>. This boundary has not been mapped at cellular resolution, but our results suggest that it coincides with the anterior *en* cells. These cells adopt the smooth fate in a *wg*-dependent manner, whereas the posterior *en* cells produce denticles (Fig. 4). We propose that at least two mechanisms instruct epidermal cell fate decisions. *wg* instructs cells to adopt the smooth fate, whereas another process restricts the *wg* signal, allowing the generation of the six denticle cell fates. □

Received 19 May; accepted 7 October 1992.

- Slack, J. M. W. *From Egg to Embryo: Regional Specification in Early Development* 2nd edn (Cambridge Univ. Press, Cambridge, 1991).
- Martinez Arias, A. & Lawrence, P. A. *Nature* **313**, 639–642 (1985).
- Rijsewijk, F. et al. *Cell* **50**, 649–657 (1987).
- van den Heuvel, M., Nusse, R., Johnston, P. & Lawrence, P. A. *Cell* **59**, 739–749 (1989).
- Lawrence, P. A. & Johnston, P. *Development* **105**, 761–767 (1989).
- Hama, C., Zehra, A. & Kornberg, T. B. *Genes Dev.* **4**, 1079–1093 (1990).
- DiNardo, S., Sher, E., Heemskerk-Jongens, J., Kassisi, J. & O'Farrell, P. H. *Nature* **332**, 604–609 (1988).
- Martinez Arias, A., Baker, N. E. & Ingham, P. W. *Development* **103**, 157–170 (1988).
- Lohs-Schardin, M., Cremer, C. & Nüsslein-Volhard, C. *Dev. Biol.* **73**, 239–255 (1979).
- Kassisi, J. A., Vansickle, E. D. & Sensabaugh, S. M. *Genetics* **128**, 751–761 (1991).
- Bejsovec, A. & Martinez Arias, A. *Development* **113**, 471–485 (1991).
- Baker, N. E. *Dev. Biol.* **125**, 96–108 (1988).
- Baker, N. E. *EMBO J.* **6**, 1765–1773 (1987).
- Jürgens, G., Wieschaus, E., Nüsslein-Volhard, C. & Kluding, H. *Willhelm Roux Arch. Dev. Biol.* **193**, 283–295 (1984).
- Hillman, R. & Lesnik, L. H. *J. Morph.* **131**, 383–396 (1970).
- Heemskerk, J., DiNardo, S., Kostriken, R. & O'Farrell, P. H. *Nature* **352**, 404–452 (1991).
- McMahon, A. P., Joyner, A. L., Bradley, A. & McMahon, J. A. *Cell* **69**, 581–595 (1992).
- González, F., Swales, L., Bejsovec, A., Skær, H. & Martinez Arias, A. *Mechanisms Dev.* **35**, 43–54 (1991).
- Nüsslein-Volhard, C., Wieschaus, E. & Kluding, H. *Willhelm Roux Arch. Dev. Biol.* **193**, 267–282 (1984).



20. Hiromi, Y., Kuroiwa, A. & Gehring, W. J. *Cell* **43**, 603–613 (1985).  
 21. Tautz, D. & Pfeifle, C. *Chromosoma* **98**, 81–85 (1989).  
 22. Campos-Ortega, J. A. & Hartenstein, V. *The Embryonic Development of Drosophila melanogaster* (Springer, Berlin, 1985).

ACKNOWLEDGEMENTS. We thank P. Ingham for the *nkx*, *Ubx* chromosome, R. White and R. Mann for *Ubx* antibody, J. Kassis for information about the *en-lacZ* stock before publication, A. Bejsovec, W. Bender and J. Heemskerk for advice and discussions and M. Boyle, C. Desplan, N. Dostatni, P. Gönczy, A. Goriely and J. Mullen for comments on the manuscript. S.D.N. is a Lucille P. Markey Scholar, and work in the lab is supported by the Markey Charitable Trust and the NIH.

## Brefeldin A inhibits Golgi membrane-catalysed exchange of guanine nucleotide onto ARF protein

Julie G. Donaldson, Dario Finazzi  
& Richard D. Klausner

Cell Biology and Metabolism Branch, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA

THE fungal metabolite brefeldin A is a powerful tool for investigating membrane traffic in eukaryotic cells<sup>1</sup>. The effects of brefeldin A on traffic are partly explained by its ability to prevent binding of cytosolic coat proteins onto membranes<sup>2–5</sup>. The non-clathrin coatomer complex<sup>6,7</sup> binds reversibly to Golgi membranes in a GTP-controlled cycle<sup>8–10</sup>. The low-molecular-mass GTP-binding protein ADP-ribosylation factor (ARF), which also associates reversibly with Golgi membranes<sup>11,12</sup>, is required for coatomer binding<sup>13</sup> and probably accounts for the control by guanine nucleotide of the coatomer–membrane interaction. Brefeldin A prevents the assembly of coatomer onto the membrane by inhibiting the GTP-dependent interaction of ARF with the Golgi membrane<sup>13</sup>, but the nature of this interaction has not been established. Here we demonstrate that Golgi membranes can specifically catalyse the exchange of GTP onto ARF and that brefeldin A prevents this function.

Myristoylated recombinant human ARF-1 (refs 14, 15) was purified and shown to bind to Golgi membranes and to be able to reconstitute the GTP-dependent, brefeldin A (BFA)-inhibitable binding of the coatomer protein  $\beta$ -COP to isolated Golgi membranes<sup>13</sup>. To measure the binding of GTP to proteins, protein-bound GTP- $\gamma$ S was trapped on nitrocellulose filters after incubation of this radioactive non-hydrolysable nucleotide with either ARF or Golgi membranes alone, or with a combination of ARF plus membranes. Either preparation on its own yielded a time-dependent increase in protein-bound [<sup>35</sup>S]GTP- $\gamma$ S which was inhibited by excess unlabelled GTP, but not by BFA (Fig. 1a, b). In contrast, when ARF and Golgi membranes were incubated together, there was an increase in the amount of protein-bound GTP- $\gamma$ S that was greater than the sum of the two components when incubated separately. This enhanced GTP- $\gamma$ S binding (the difference between the observed and expected sum) was inhibited by inclusion of BFA in the reaction (Fig. 1c). Half-maximal inhibition of the enhanced GTP- $\gamma$ S binding by BFA was at  $\sim 10 \mu\text{M}$  BFA (Fig. 2a). Although the enhanced binding of GTP- $\gamma$ S increased with ARF concentration and saturated at  $25 \mu\text{g ml}^{-1}$ , BFA was quantitatively as effective at saturating concentrations of ARF (Fig. 2b).

The use of the non-hydrolysable GTP analogue GTP- $\gamma$ S ruled out the possibility that BFA inhibited the interaction of ARF with membranes by stimulating the hydrolysis of GTP. We next examined guanine-nucleotide binding using either [ $\alpha$ -<sup>32</sup>P]- or [ $\gamma$ -<sup>32</sup>P]GTP. If bound GTP is hydrolysed, then only the  $\alpha$ -labelled compound will be detected as the  $\gamma$ -phosphate will be

lost on hydrolysis of GTP to GDP. Figure 3a shows the amount of protein-bound guanine nucleotide obtained after incubation of ARF and Golgi membranes with  $\alpha$ -labelled nucleotide: as with GTP- $\gamma$ S, binding of the hydrolysable  $\alpha$ -labelled GTP increases with time and is halted by addition of BFA at 20 min into the incubation. These results contrast with those obtained with  $\gamma$ -labelled phosphate, which gave no increase in binding when coincubated with ARF and Golgi membranes (Fig. 3b). With ARF alone no difference in the level of binding was

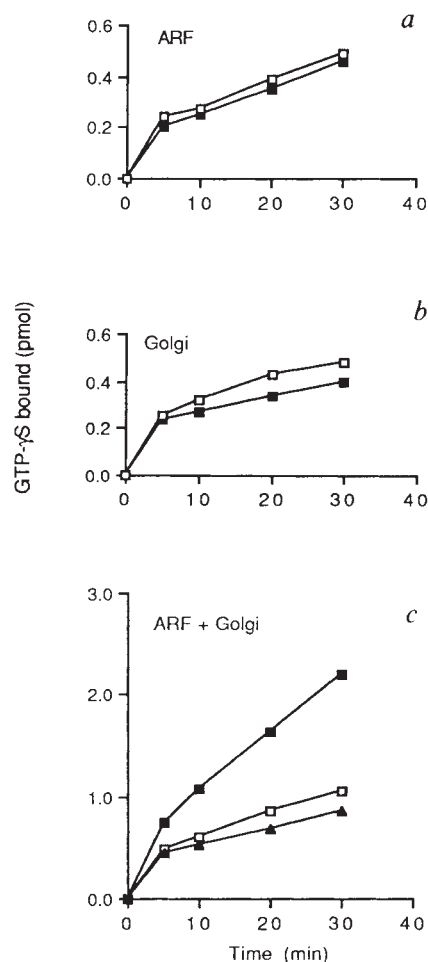
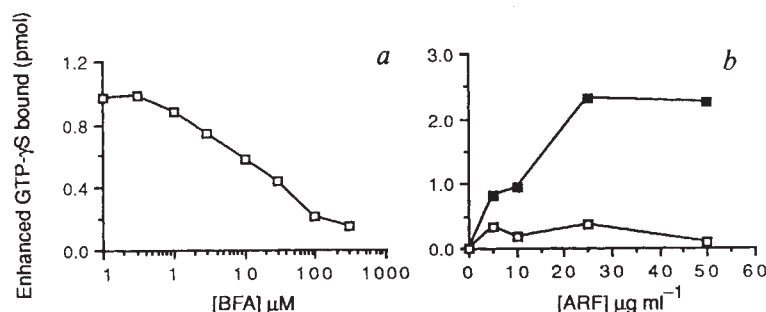


FIG. 1. Time course of GTP- $\gamma$ S binding to ARF and Golgi membranes. Myristoylated ARF (a), Golgi membranes (b), or both (c) were incubated in buffer containing  $1.0 \mu\text{M}$  [<sup>35</sup>S]GTP- $\gamma$ S in the absence (solid squares) or presence (open squares) of BFA ( $200 \mu\text{M}$ ) for the indicated times and then the amount of protein-bound GTP- $\gamma$ S (pmol) was measured. In c, the sum of the GTP- $\gamma$ S bound to ARF and Golgi membrane from a and b is also plotted (solid triangles). The difference between the observed binding with ARF and Golgi membrane incubations and the expected sum represents the enhanced GTP- $\gamma$ S binding.

METHODS. Recombinant myristoylated ARF was purified from *Escherichia coli* coexpressing the human ARF-1 gene and *N*-myristoyltransferase<sup>14,15</sup>. A Golgi-enriched membrane fraction from Chinese hamster ovary cells was obtained by sucrose gradient centrifugation as described<sup>25</sup>. Incubations were carried out in a  $100\text{-}\mu\text{l}$  reaction mixture containing  $25 \text{ mM}$  HEPES-KOH, pH 7.0,  $25 \text{ mM}$  KCl,  $2.5 \text{ mM}$  MgCl<sub>2</sub>,  $0.2 \text{ M}$  sucrose,  $1 \text{ mM}$  DTT,  $150 \mu\text{g}$  BSA,  $1 \text{ mM}$  ATP,  $2\text{--}5 \mu\text{g}$  recombinant ARF,  $4\text{--}6 \mu\text{g}$  Golgi membranes and [<sup>35</sup>S]GTP- $\gamma$ S ( $1 \mu\text{M}$ ,  $0.7\text{--}1.0 \text{ Ci mmol}^{-1}$ ) at  $30^\circ\text{C}$  for the indicated times. The reaction was terminated by the addition of  $2 \text{ ml}$  ice-cold  $10 \text{ mM}$  HEPES-KOH, pH 7.0, and the amount of protein-bound nucleotide determined by filtration through BA-85 nitrocellulose filters followed by five  $2\text{-ml}$  washes with cold buffer<sup>26</sup>. Nonspecific binding was measured in a separate set of samples containing  $1.0 \text{ mM}$  GTP and this value was subtracted from the binding measured in the absence of added GTP. Results shown are representative of three independent experiments.

FIG. 2 Effect of BFA and ARF concentrations on the enhanced binding of GTP- $\gamma$ S. Golgi membranes and ARF were incubated for 20 min before measuring the protein-bound GTP- $\gamma$ S (pmol) as described in Fig. 1 legend. *a*, Incubation in the presence of increasing concentrations of BFA. *b*, Incubation of Golgi membrane (6  $\mu$ g) with increasing concentrations of ARF in the absence (solid squares) or presence (open squares) of 200  $\mu$ M BFA.

METHODS. Incubations and measurement of binding were carried out as for Fig. 1. The sum of the protein-bound GTP- $\gamma$ S following incubation of ARF and Golgi membranes alone in *a* and *b* was determined and subtracted from that observed with co-incubation of ARF and membrane to yield the enhanced binding of GTP- $\gamma$ S, which is shown in pmol. Data are representative of two independent experiments.



observed between [ $\gamma$ -<sup>32</sup>P]- and [ $\alpha$ -<sup>32</sup>P]GTP, which is consistent with the fact that ARF lacks GTPase activity<sup>15</sup>. These results are best explained if all of the enhanced GTP binding that takes place when ARF and membranes are mixed is accompanied by rapid hydrolysis of the GTP.

Current models for the interaction between ARF and target membranes propose that ARF-GTP is associated with Golgi membranes, while ARF-GDP is present in the cytosol<sup>10,12</sup>. Recombinant ARF does indeed bind stably to Golgi-enriched membranes in the presence of GTP- $\gamma$ S, but not in the presence of GTP<sup>10,12</sup>, presumably because the GTP is rapidly hydrolysed. This model predicts that in the presence of [ $\alpha$ -<sup>32</sup>P]GTP all of the ARF-bound nucleotide should be in the supernatant after the Golgi membranes are pelleted. Figure 3c shows that all of the enhanced protein-bound nucleotide is indeed in the supernatant and not with the Golgi membrane when [ $\alpha$ -<sup>32</sup>P]GTP is used in the assay. This is in contrast to results obtained with GTP- $\gamma$ S, when much of the labelled nucleotide is found in the Golgi membrane pellet (not shown). Evidence that the protein-bound nucleotide in the supernatant was bound to ARF came from the co-migration of the bound nucleotide with recombinant ARF in gel filtration chromatography (not shown), although we cannot exclude the possibility that ARF induces nucleotide binding and the release of a Golgi membrane GTP-binding protein. The complete inhibition of nucleotide binding to ARF by BFA focuses the action of this drug on a Golgi membrane-induced guanine nucleotide-exchange reaction.

We next examined the specificity of this membrane-induced, guanine-nucleotide binding to ARF and its sensitivity to BFA. Trypsin treatment of Golgi membranes abolishes the binding of ARF to Golgi membranes in the presence of GTP- $\gamma$ S<sup>12</sup>. Preincubation of the membranes with trypsin significantly inhibited binding of [<sup>32</sup>P]GTP to Golgi membranes alone and abolished 85% of the membrane-induced, enhanced binding (Fig. 4a). Furthermore, when rod outer segment membranes (which do not support ARF-dependent coatamer binding *in vitro*) were used, no enhanced or BFA-sensitive exchange was observed (data not shown). Incubation of purified ARF with a mixture of dimyristoyl phosphatidylcholine and sodium cholate allows membrane-independent exchange of GTP onto ARF<sup>16</sup>, and we have found that exchange of GTP onto ARF under these conditions is independent of the presence of BFA (not shown), suggesting that the effect of the drug is specific for the protein-dependent, Golgi-induced exchange activity. We have also tested several inactive analogues of BFA which have no effect on the Golgi complex or on the distribution of ARF and  $\beta$ -COP in intact cells (not shown); these inactive analogues, including B-31, a stereoisomer at the C7 hydroxyl, have no effect on the enhanced exchange reaction in this assay (Fig. 4b).

ARF is probably involved in several membrane traffic events in the cell<sup>17-20</sup>. As it is necessary for the assembly of cytosolic coat proteins onto Golgi membranes and because it is the ARF-membrane interaction step that is inhibited by BFA, we have investigated the mechanism of action of this drug further.

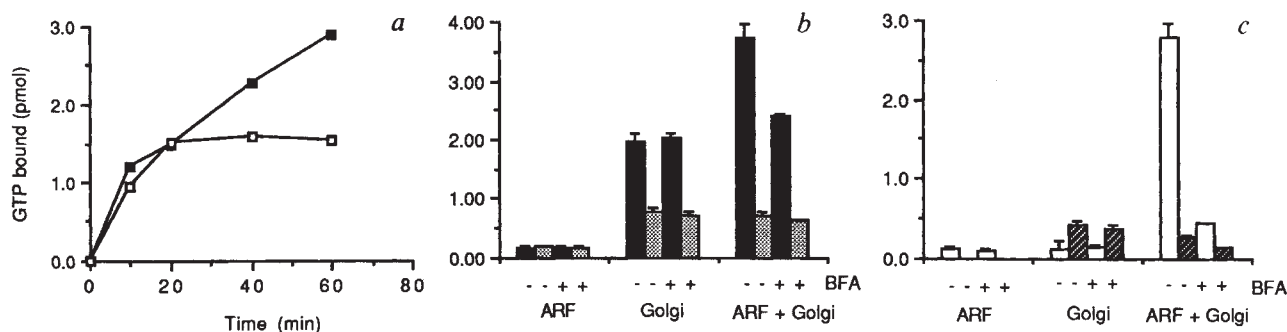


FIG. 3 Binding of [ $\alpha$ -<sup>32</sup>P]GTP and [ $\gamma$ -<sup>32</sup>P]GTP to ARF and Golgi membranes. *a*, Time course of total binding of [ $\alpha$ -<sup>32</sup>P]GTP to incubations containing ARF and Golgi membranes (filled squares). After 20 min BFA (300  $\mu$ M) was added to some samples (empty squares). *b*, Comparison of total binding of [ $\alpha$ -<sup>32</sup>P]GTP (solid bars) to ARF alone, Golgi membranes alone, or ARF plus Golgi membranes, with that of [ $\gamma$ -<sup>32</sup>P]GTP (stippled bars) in the absence and presence of BFA (300  $\mu$ M) after 30 min incubation. *c*, Distribution of protein-bound nucleotide after incubation of ARF, Golgi membranes or ARF plus Golgi membrane with [ $\alpha$ -<sup>32</sup>P]GTP for 30 min in the absence and presence of BFA (300  $\mu$ M), followed by pelleting of the membranes and measurement

of filter-trapped nucleotide in the supernatant (open bars) and membrane fractions (hatched bars).

METHODS. Incubations were carried out as for Fig. 1 with either [ $\alpha$ -<sup>32</sup>P]GTP or [ $\gamma$ -<sup>32</sup>P]GTP (1.0  $\mu$ M, 1.0 Ci m mol<sup>-1</sup>). For *c*, after 30 min incubation membranes were pelleted at 14,000g for 10 min (in a Beckman TLA100.2 rotor) and the protein-bound nucleotide in the supernatant and in the resuspended pellet were estimated by filter binding assay. Total pmol of bound  $\alpha$ - or  $\gamma$ -labelled nucleotide is shown for these three experiments, which have been repeated once. For *b* and *c*, the mean and standard error for triplicate samples is shown.



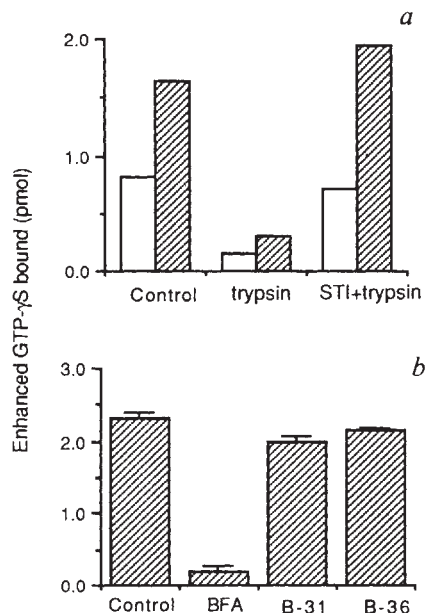


FIG. 4 Effect of protease treatment of Golgi membranes and BFA analogues on [ $^{32}$ P]GTP binding. *a*, Golgi membranes were untreated (control), treated at 4 °C with 30  $\mu$ g ml $^{-1}$  trypsin for 60 min followed by addition of soybean trypsin inhibitor, or treated with a mixture of 30  $\mu$ g ml $^{-1}$  trypsin and 300  $\mu$ g ml $^{-1}$  soybean trypsin inhibitor (STI+trypsin) at 4 °C for 60 min before incubation with [ $\alpha$ - $^{32}$ P]GTP in the absence (open bars) or presence (hatched bars) of ARF. In the absence of ARF, total protein-bound nucleotide is shown; in the presence of ARF, the enhanced binding of nucleotide (pmol) is shown. *b*, ARF plus Golgi membranes were incubated with [ $\alpha$ - $^{32}$ P]GTP in the presence of BFA, or analogues of BFA, B-31 or B-36 (300  $\mu$ M each) for 30 min before assessment of bound nucleotide. The enhanced binding of nucleotide is shown with the standard error for triplicate incubations.

We have shown that there is a protease- and BFA-sensitive, Golgi-associated factor that induces guanine nucleotide binding to ARF. We presume that this enhanced binding represents the action of an exchange protein, based on the likelihood that the ARF in the preparation has bound nucleotide<sup>14</sup>. It has become clear that exchange of guanine nucleotides onto low-molecular-mass GTP-binding proteins is a complex process affected by different stimulatory and inhibitory proteins<sup>21–24</sup>. It remains to be determined whether BFA acts directly on a GTP-exchange protein or indirectly through a modulating factor, and whether other exchange proteins are affected by BFA. A drug that specifically interferes with this critical step in the activation of these GTP-binding proteins may have pharmacological potential for manipulating processes involving such exchange activity. □

Received 5 August; accepted 9 October 1992.

- Klausner, R. D., Donaldson, J. G. & Lippincott-Schwartz, J. *J. Cell Biol.* **116**, 1071–1080 (1992).
- Donaldson, J. G., Lippincott-Schwartz, J., Bloom, G. S., Kreis, T. E. & Klausner, R. D. *J. Cell Biol.* **111**, 2295–2306 (1990).
- Orci, L. *et al. Cell* **64**, 1183–1195 (1991).
- Robinson, M. S. & Kreis, T. E. *Cell* **69**, 129–138 (1992).
- Narula, N. *et al. J. Cell Biol.* **117**, 27–38 (1992).
- Serafini, T. *et al. Nature* **349**, 215–220 (1991).
- Waters, M. G., Serafini, T. & Rothman, J. E. *Nature* **349**, 248–251 (1991).
- Melançon, P. *et al. Cell* **51**, 1053–1062 (1987).
- Donaldson, J. G., Lippincott-Schwartz, J. & Klausner, R. D. *J. Cell Biol.* **112**, 579–588 (1991).
- Donaldson, J. G., Kahn, R. A., Lippincott-Schwartz, J. & Klausner, R. D. *Science* **254**, 1197–1199 (1991).
- Stearns, T., Willingham, M. C., Botstein, D. & Kahn, R. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1238–1242 (1990).
- Serafini, T. *et al. Cell* **67**, 239–254 (1991).
- Donaldson, J. G., Cassel, D., Kahn, R. A. & Klausner, R. D. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6408–6412 (1992).
- Weiss, O., Holden, J., Rulka, C. & Kahn, R. A. *J. Biol. Chem.* **264**, 21066–21072 (1989).
- Duronio, R. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **87**, 1506–1510 (1990).
- Kahn, R. A. & Gilman, A. G. *J. Biol. Chem.* **261**, 7906–7911 (1986).
- Balch, W. E., Kahn, R. A. & Schwaninger, R. *J. Biol. Chem.* **267**, 13053–13061 (1992).
- Kahn, R. A. *et al. J. Biol. Chem.* **267**, 13039–13046 (1992).
- Lenhard, J. M., Kahn, R. A. & Stahl, P. D. *J. Biol. Chem.* **267**, 13047–13052 (1992).

- Taylor, T. C., Kahn, R. A. & Melançon, P. *Cell* **70**, 69079 (1992).
- Bourne, H. R., Sanders, D. A. & McCormick, F. *Nature* **348**, 125–131 (1990).
- Hall, A. *Science* **249**, 635–640 (1990).
- Kikuchi, A. *et al. J. Biol. Chem.* **267**, 14611–14615 (1992).
- Shou, C., Farnsworth, C. L., Neel, B. G. & Feig, L. A. *Nature* **358**, 351–354 (1992).
- Balch, W. E., Glick, B. S. & Rothman, J. E. *Cell* **39**, 525–536 (1984).
- Kahn, R. A. *Meth. Enzym.* **195**, 233–243 (1991).

ACKNOWLEDGEMENTS. We thank P. Peters, D. Cassel and all our colleagues in the CBMB for comments and suggestions, and C. Yim for preparation of Golgi membranes. D.F. was supported by a fellowship from AIRC, the Italian Association for Cancer Research.

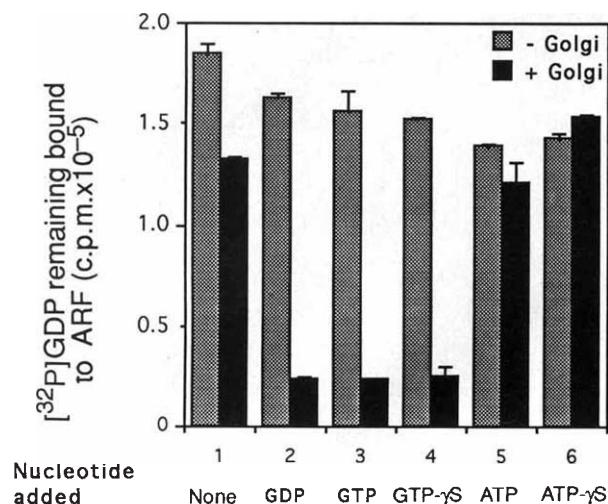
## Inhibition by brefeldin A of a Golgi membrane enzyme that catalyses exchange of guanine nucleotide bound to ARF

J. Bernd Helms & James E. Rothman

Program in Cellular Biochemistry and Biophysics, Rockefeller Research Laboratory, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA

A WIDE variety of membrane transformations important in intracellular transport are inhibited by the fungal metabolite brefeldin A (refs 1–4), implying that the target for this drug is central to the formation and maintenance of subcellular compartments<sup>5</sup>. Brefeldin A added to cells causes the rapid and reversible dissociation of a Golgi-associated peripheral membrane protein ( $M_r$  110,000)<sup>6</sup> which was found to be identical to one of the subunits of the coat of Golgi-derived (non-clathrin) coated vesicles,  $\beta$ -COP<sup>7,8</sup>, implying that brefeldin A prevents transport by blocking the assembly of coats and thus the budding of enclosed vesicles<sup>9</sup>. In addition to the coatmer (a cytosol-derived complex of seven polypeptide chains, one of which is  $\beta$ -COP<sup>10</sup>), the non-clathrin (COP) coat of Golgi-derived vesicles contains stoichiometric amounts of a small ( $M_r$  ~20,000) GTP-binding protein, the ADP-ribosylation factor (ARF)<sup>11</sup>. Binding of ARF to Golgi membranes is necessary before coatmer/ $\beta$ -COP can bind these membranes (ref. 12; and D. J. Palmer *et al.*, manuscript submitted), so the primary effect of brefeldin A seems to be on the reaction responsible for ARF binding. Indeed, like  $\beta$ -COP, ARF is dissociated from the Golgi complex by treatment with brefeldin A and brefeldin A prevents ARF from associating *in vitro*<sup>13</sup>, but the mechanism of this action by brefeldin A has been unclear. Here we report the discovery of an enzyme in a Golgi-enriched fraction that catalyses guanine nucleotide (GDP-GTP) exchange on ARF-1 protein, and which is inhibited by brefeldin A. We suggest that activation of ARF proteins for membrane localization by compartmentalized exchange enzymes is in general the first committed step in membrane transformation pathways.

ARF purified from cytosol as well as recombinant ARF purified from bacteria is in the GDP form<sup>14,15</sup>. Recombinant *N*-myristoylated ARF was purified from an *Escherichia coli* strain expressing both bovine ARF-1 and yeast *N*-myristoyl transferase using established procedures<sup>15,16</sup>, and preloaded with [ $\alpha$ - $^{32}$ P]GDP. To test for a nucleotide-exchanging activity in Golgi membranes, ARF-[ $^{32}$ P]GDP was incubated with Golgi membranes and the radioactive GDP remaining bound was assayed as radioactivity bound to a nitrocellulose filter<sup>17</sup>. Figure 1 shows that [ $^{32}$ P]GDP remained stably associated with ARF unless both Golgi membranes and excess guanine nucleotides (GDP, GTP or GTP- $\gamma$ S) were present. This effect is specific for guanine nucleotides as ATP or ATP- $\gamma$ S could not replace guanine nucleotides in promoting release of bound GDP (Fig. 1). Evidently, the rate of spontaneous nucleotide exchange/release is very low (Fig. 1). The exchange activity is protease-sensitive and remains associated with Golgi membranes after



extraction with 1 M KCl, which is indicative of Golgi-bound protein-catalysed reaction (data not shown).

This experiment demonstrates a guanine nucleotide-dependent release of GDP from ARF that requires Golgi membranes. To demonstrate that a guanine nucleotide (GTP-γS) is in fact exchanged for GDP bound to ARF, purified myristoylated ARF protein (with unlabelled GDP bound) was incubated under the same conditions but with GTP-γS labelled with <sup>35</sup>S, and the amount of radioactivity bound to protein determined on nitrocellulose filters (Fig. 2). Binding of GTP-γS required both ARF protein and Golgi membranes. The activity in Golgi membranes was, as expected, sensitive to proteinase K and resistant to salt extraction (Fig. 2). Photocrosslinking of the bound [<sup>35</sup>S]GTP-γS (ref. 18) and SDS-PAGE confirmed that the filter-bound radioactivity was in ARF (data not shown). Comparing the data in Figs 1 and 2, about 3.3 pmol of GTP-γS enters ARF for every 2.7 pmol of GDP that leaves, consistent with a 1-for-1 exchange mechanism. Bovine brain cytosol tested at the same protein concentration as Golgi membranes did not stimulate nucleotide exchange (Fig. 2), suggesting that the salt-resistant

FIG. 2 Incorporation of [<sup>35</sup>S]GTP-γS in ARF catalysed by Golgi membranes. ARF (bar 1), rat liver Golgi membranes (bar 2), or ARF and Golgi membranes (bars 3–7) were incubated with [<sup>35</sup>S]GTP-γS and the radiolabelled nucleotide bound to ARF protein was measured as described in Fig. 1. Bar 4, mock digestion of Golgi membranes with proteinase K in the presence of 2 mM PMSF; bar 5, pre-treatment of Golgi membranes with proteinase K; bar 6, salt-washed rat liver Golgi membranes and ARF; bar 7, carbonate-treated rat liver Golgi membranes and ARF; bar 8, bovine brain cytosol and ARF. METHODS. Incubations were for 15 min at 37 °C in 25 mM HEPES-KOH, pH 7.2, 20 mM KCl, 2.5 mM magnesium acetate, ovalbumin (1.6 mg ml<sup>-1</sup>, 0.2 M sucrose and 5 μM [<sup>35</sup>S]GTP-γS (2 × 10<sup>4</sup> c.p.m. pmol<sup>-1</sup>) in a total volume of 50 μl. Other components were added as indicated: ARF (3 μg), rat liver Golgi membranes (2.7 μg), salt-washed rat liver Golgi membranes (1.7 μg) and carbonate treated membranes (1 μg). Golgi membranes were pre-incubated with proteinase K (50 μg ml<sup>-1</sup>) for 5 min at 20 °C and subsequently with 2 mM PMSF for 5 min at 20 °C. Mock digestion of Golgi membranes was done in the presence of both proteinase K (50 μg ml<sup>-1</sup>) and PMSF (2 mM) for 10 min at 20 °C. Membranes were stripped of peripherally associated proteins by 1 M KCl (salt-washed) or 0.1 M Na<sub>2</sub>CO<sub>3</sub> (carbonate-treated) as described<sup>30</sup>, except that extraction was at 0 °C. Bovine brain cytosol was prepared as in ref. 31.

FIG. 1 Effect of nucleotides on the release of GDP[α-<sup>32</sup>P] bound to ARF. Pre-loaded ARF[α-<sup>32</sup>P]GDP was incubated in the absence (open bars) or presence (filled bars) of Golgi membranes and unlabelled nucleotides (bars 1–6). The release of [<sup>32</sup>P]GDP bound to ARF was measured. No nucleotide (bar 1), GDP (bar 2), GTP (bar 3), GTP-γS (bar 4), ATP (bar 5) and ATP-γS (bar 6).

METHODS. The release assay contained 3 μg of ARF-[α-<sup>32</sup>P]GDP (6 × 10<sup>4</sup> c.p.m. pmol<sup>-1</sup>), 25 mM HEPES-KOH, pH 7.2, 2.5 mM magnesium acetate, 20 mM KCl, 0.2 M sucrose, 1.6 mg ml<sup>-1</sup> ovalbumin, 50 μM unlabelled nucleotide in a total volume of 50 μl and in the absence or presence of rat liver Golgi membranes (8.1 μg). After incubation for 15 min at 37 °C, the release of the ARF bound [α-<sup>32</sup>P]GDP was determined by nitrocellulose filter-binding assay<sup>17</sup>. Recombinant myristoylated ARF was overexpressed and purified to homogeneity as described<sup>15,16</sup>. Rat liver Golgi membrane was prepared according to ref. 30. Recombinant myristoylated ARF was preloaded with [α-<sup>32</sup>P]GDP by incubating 130 μg ARF, 40 μg rat liver Golgi membranes and 3 μM [α-<sup>32</sup>P]GTP (6 × 10<sup>4</sup> c.p.m. pmol<sup>-1</sup>) in a total volume of 500 μl of buffer as described for the release assay for 15 min at 37 °C. Radiolabelled ARF recovered from the supernatant was all in the GDP form (J.B.H., manuscript submitted). After incubation, ARF-[α-<sup>32</sup>P]GDP was desalted on a 4-ml Sephadex G-25 column equilibrated with 25 mM HEPES-KOH, pH 7.2, 2.5 mM magnesium acetate, 50 mM KCl and 0.2 M sucrose. The ARF-containing fractions were pooled and concentrated by ultracentrifugation (Amicon 10 membrane) to 500 μl.

exchange activity is not cytosol-derived. But because alkali extraction (0.1 M sodium carbonate, pH 11) prevented Golgi membranes from catalysing the nucleotide-exchange reaction, it cannot be firmly concluded that the exchange activity requires an integral membrane protein.

Brefeldin A (BFA) profoundly inhibited the membrane-dependent nucleotide-exchange reaction (Fig. 3a). As expected<sup>13</sup>, BFA also inhibited the binding of myristoylated ARF to Golgi membranes (Fig. 3b). The specificity of inhibition is shown by the fact that an analogue of BFA, in which the hydroxyl groups at positions 4 and 7 are modified (from M. L. Snapper and S. Schreiber), did not inhibit these reactions.

It has been proposed that ARF proteins are activated by direct phosphorylation of bound GDP (by nucleoside diphosphate kinase (NDK)) to yield bound GTP without release of the bound GDP<sup>19</sup>, rather than by nucleotide exchange, but this conclusion has since been withdrawn<sup>20</sup>. Our results showing that radioactivity from [<sup>35</sup>S]GTP-γS is transferred to ARF (Fig. 2) could in principle be explained by the NDK mechanism. But our data demonstrating a concomitant release of GDP rules this out as

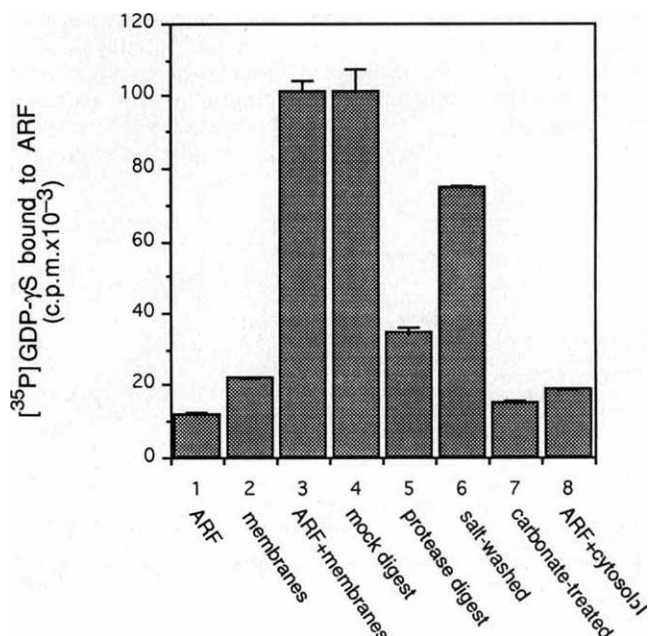
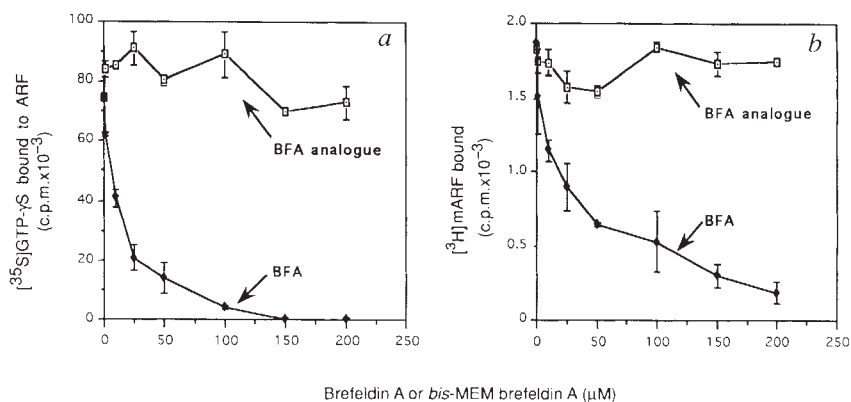




FIG. 3 Inhibition of both *a*, nucleotide-exchange activity and *b*, membrane binding of ARF by brefeldin A. Rat liver Golgi membranes were preincubated for 10 min at 37 °C with different concentrations of BFA (diamonds) or BFA analogue (bis-MEM BFA) (squares). For determining the effect on the nucleotide-exchange enzyme, ARF and [<sup>35</sup>S]GTP-γS were added and the samples were incubated for 15 min at 37 °C and the ARF-bound [<sup>35</sup>S]GTP-γS was determined as described for Fig. 2. For the effect of BFA on the binding of ARF to Golgi membranes, [<sup>3</sup>H]myristate-labelled ARF ([<sup>3</sup>H]mARF) and unlabelled GTP-γS were added and samples were incubated for 20 min at 37 °C. Golgi membranes were then pelleted through a sucrose cushion and the radioactivity associated with the membranes counted.

**METHODS.** *a*, After pretreatment of rat liver Golgi membranes with BFA or BFA analogue (4,7-bis(methoxyethoxymethyl)BFA, abbreviated as bis-MEM-BFA), Golgi membranes (2.7 μg) were incubated with ARF (3 μg) and [<sup>35</sup>S]GTP-γS (2 × 10<sup>4</sup> c.p.m. pmol<sup>-1</sup>) as described in Fig. 2 legend. BFA and bis-MEM-BFA were solubilized in methanol (10 mM), the final methanol concentration in all incubations being 2% (v/v). Data are corrected for spontaneous incorporation of [<sup>35</sup>S]GTP-γS into Golgi membranes (15,218 c.p.m.) and ARF protein (16,924 c.p.m.). *b*, Pretreated Golgi membranes were incubated with [<sup>3</sup>H]mARF (4.4 × 10<sup>4</sup> c.p.m.) and unlabelled GTP-γS (50 μM) and other components as described for the nucleotide-



exchange assay (Fig. 1 legend) in a total volume of 50 μl. After incubation, 45 μl of the mixture was loaded onto 165 μl 25% sucrose (w/v) and centrifuged for 30 min at 14,000 r.p.m. at 4 °C in an Eppendorf centrifuge. The pellet was resuspended and counted for radioactivity. Data are corrected for nonspecific (protease-insensitive) binding of ARF to Golgi membranes (604 c.p.m.). [<sup>3</sup>H]mARF was synthesized by addition of [<sup>3</sup>H]myristic acid during overexpression of ARF in *E. coli*. After purification ARF protein is the only protein labelled with [<sup>3</sup>H]myristate, as confirmed by SDS-PAGE and autoradiography (not shown).

the principal mechanism under our conditions, as does the fact that GDP release is not stimulated by ATP or ATP-γS, both of which are substrates for NDK. Moreover, ARF binding to membranes, the result of ARF activation, is promoted by GMP-PNP as well as by GTP-γS but not by ATP-γS (J.B.H., manuscript submitted), ruling out an NDK mechanism here too. The fact that nucleotide exchange is inhibited by BFA provides a strong pharmacological argument that underscores the physiological significance of the nucleotide exchange mechanism for ARF activation.

The simplest interpretation of our results is that BFA works in the Golgi by blocking GTP-GDP exchange in ARF, thereby blocking the activation of ARF for membrane binding and coat assembly, thus preventing budding of transport vesicles. BFA may act directly on the nucleotide-exchanging enzyme by binding to it, but this cannot be demonstrated unless purified enzyme becomes available. It remains to be verified that the exchange enzyme for ARF-1 is localized on Golgi membranes rather than being associated with a trace contaminant, but this is most unlikely in view of the established effects of BFA on the Golgi apparatus. The compartmental localization of the exchange enzyme, which appears to be tightly bound to membranes (that is, it is resistant to 1 M KCl) (Fig. 2), could be essential for

determining the localization of its product, membrane-bound ARF-GTP, and consequently the sites for vesicle budding. Because myristoylated ARF-GTP, but not ARF-GDP, can insert spontaneously into liposomes<sup>21</sup>, the localization of exchanging enzymes may even be the sole basis for localizing ARF proteins<sup>11</sup>.

Animal cells contain a family of related ARF proteins<sup>22</sup>, and one or more of its members is important for transport to the Golgi from the endoplasmic reticulum<sup>23,24</sup>, for intercisternal transport in the Golgi<sup>25</sup>, for endocytosis and related events<sup>26</sup>, and for nuclear envelope assembly following mitosis<sup>27</sup>, although it is not yet clear whether particular family members are specialized for particular membrane transformations. There may well be a family of differentially localized (and typically BFA-sensitive) nucleotide-exchanging enzyme specific for particular ARF proteins that would impart compartmental specificity to membrane vesicle budding and fusion events<sup>28</sup>. It has been shown that BFA also inhibits the binding of clathrin coat protein γ-adaptin to membranes in permeabilized cells<sup>29</sup>. Thus the binding of ARF could be the first committed step not only in the assembly of non-clathrin coated vesicles but also of other types of coated vesicles as well. □

Received 17 August; accepted 9 October 1992.

1. Fujiwara, T., Oda, K., Yokota, S., Takatsuki, A. & Ikehara, Y. *J. Biol. Chem.* **263**, 18545-18552 (1988).
2. Wood, S. A., Park, J. E. & Brown, W. J. *Cell* **67**, 591-600 (1991).
3. Lippincott-Schwartz, J. et al. *Cell* **67**, 601-616 (1991).
4. Hunziker, W., Whitney, A. & Mellman, I. *Cell* **67**, 617-627 (1991).
5. Pelham, H. R. B. *Cell* **67**, 449-451 (1991).
6. Donaldson, J. G., Lippincott-Schwartz, J., Bloom, G. S., Kreis, T. E. & Klausner, R. D. *J. Cell Biol.* **111**, 2295-2306.
7. Allan, V. J. & Kreis, T. E. *J. Cell Biol.* **103**, 2229-2239 (1986).
8. Serafini, T. et al. *Nature* **349**, 215-220 (1991).
9. Orci, L. et al. *Cell* **64**, 1183-1195 (1991).
10. Waters, M. G., Serafini, T. & Rothman, J. E. *Nature* **349**, 248-251 (1991).
11. Serafini, T. et al. *Cell* **67**, 239-253 (1991).
12. Donaldson, J. G., Cassel, D., Kahn, R. A. & Klausner, R. D. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6408-6412 (1992).
13. Donaldson, J. G., Kahn, R. A., Lippincott-Schwartz, J. & Klausner, R. D. *Science* **254**, 1197-1199 (1991).
14. Kahn, R. A. & Gilman, A. G. *J. Biol. Chem.* **261**, 7906-7911 (1986).
15. Weiss, O., Holden, J., Rulka, C. & Kahn, R. A. *J. Biol. Chem.* **264**, 21066-21072 (1989).
16. Duronio, R. J. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1506-1510 (1990).

17. Northup, J. K., Smigel, M. D. & Gilman, A. G. *J. Biol. Chem.* **257**, 11416-11423 (1982).
18. Schleicher, M., Witke, W. & Isenberg, G. *FEBS Lett.* **200**, 156-160.
19. Randazzo, P. A., Northup, J. K. & Kahn, R. A. *Science* **254**, 850-853 (1991).
20. Randazzo, P. A., Kahn, R. A. & Northup, J. K. *Science* **257**, 862 (1992).
21. Kahn, R. A. *J. Biol. Chem.* **266**, 15595-15597 (1991).
22. Tsuchiya, M., Price, S. R., Tsai, S.-C., Moss, J. & Vaughan, M. *J. Biol. Chem.* **266**, 2772-2777 (1991).
23. Stearns, T., Willingham, M. C., Botstein, D. & Kahn, R. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1238-1242 (1990).
24. Balch, W. E., Kahn, R. A. & Schwaninger, R. *J. Biol. Chem.* **267**, 13053-13061 (1992).
25. Kahn, R. A. et al. *J. Biol. Chem.* **267**, 13039-13046 (1992).
26. Lenhard, J. M., Kahn, R. A. & Stahl, P. D. *J. Biol. Chem.* **267**, 13047-13052 (1992).
27. Boman, A. L., Taylor, T. C., Melançon, P. & Wilson, K. L. *Nature* **358**, 512-514 (1992).
28. Rothman, J. E. & Orci, L. *Nature* **355**, 409-415 (1992).
29. Robinson, M. S. & Kreis, T. E. *Cell* **69**, 129-138 (1992).
30. Clary, D. O. & Rothman, J. E. *J. Biol. Chem.* **265**, 10109-10117 (1990).
31. Malhotra, V., Serafini, T., Orci, L., Shepherd, J. C. & Rothman, J. E. *Cell* **58**, 329-336 (1989).

**ACKNOWLEDGEMENTS.** We thank M. L. Snapper and S. Schreiber for the BFA analogue, and J. I. Gordon and R. A. Kahn for plasmids pBB131 (N-myristoyltransferase) and pOW12 (ARF) respectively. This work was supported by an NIH grant to J.E.R. and a fellowship from the Netherlands Organization of Scientific Research (NWO) to J.B.H.

# Ornithine decarboxylase activity is critical for cell transformation

Merja Auvinen, Aino Paasinen, Leif C. Andersson & Erkki Hölttä\*

Department of Pathology, University of Helsinki, Haartmaninkatu 3 SF-00290 Helsinki, Finland

\* To whom correspondence should be addressed

**THE enzyme ornithine decarboxylase is the key regulator of the synthesis of polyamines<sup>1-3</sup> which are essential for cell proliferation<sup>4,5</sup>. Expression of this enzyme is transiently increased upon stimulation by growth factors<sup>1-3</sup>, but becomes constitutively activated during cell transformation induced by carcinogens<sup>6,7</sup>, viruses<sup>8-10</sup> or oncogenes<sup>11-14</sup>. To test whether ornithine decarboxylase could be a common mediator of transformation and oncogenic itself, we transfected NIH3T3 cells with expression vectors carrying the complementary DNA encoding human ornithine decarboxylase in sense and antisense orientations. The increased expression of the enzyme (50–100-times endogenous levels) induced not only cell transformation, but also anchorage-independent growth in soft agar and increased tyrosine phosphorylation of a protein of *M*<sub>r</sub> 130K. Expression of ornithine decarboxylase antisense RNA was associated with an epithelioid morphology and reduced cell proliferation. Moreover, blocking the endogenous enzyme using specific inhibitor or synthesizing antisense RNA prevented transformation of rat fibroblasts by temperature-sensitive *v-src* oncogene. Our results imply that the gene encoding ornithine decarboxylase is a proto-oncogene central for regulation of cell growth and transformation.**

The activity of ornithine decarboxylase (ODC) is strictly regulated during the normal cell cycle: it peaks in G1 before the onset of DNA synthesis and rises again in G2/M eliciting comparable fluctuations in polyamine synthesis. The functions of polyamines include stimulation of DNA, RNA and protein synthesis and stabilization of membrane and cytoskeletal structures<sup>1-3,15,16</sup>. Given that cell transformation in general is associ-

ated with a large increase in ODC, we reasoned that this enzyme might be critical for malignant transformation. We therefore inserted human ODC cDNA into three expression vectors with promoters of different strength and transfected NIH3T3 cells with these constructs and with control plasmids and a selectable marker (pSV2neo<sup>17</sup>) if not present in the vector. Neomycin (G418)-resistant cells were isolated after two weeks of selection and their growth properties investigated. To avoid problems inherent in clonal variation, data on uncloned cultures are presented.

Transfection of NIH3T3 cells with a pLTRpoly/ODC construct elicited a complete morphological transformation (Fig. 1) in two separate experiments. The cells grew in a criss-cross fashion with no contact inhibition, and formed multiple foci in monolayer cultures (Fig. 1b). They showed disintegration of their actin filaments (results not shown) and had acquired the ability to grow in soft agar, forming large colonies (Fig. 1e) almost as efficiently as the reference E4 cells transformed by the oncogene *c-Ha-ras*<sup>Val12</sup> (ref. 11). In contrast, cells expressing the ODC antisense construct showed an epithelioid morphology, were inhibited by increased density of cells and did not form foci. They did not proliferate in soft agar either, but remained as single cells like the untransfected cells (Fig. 1c, f).

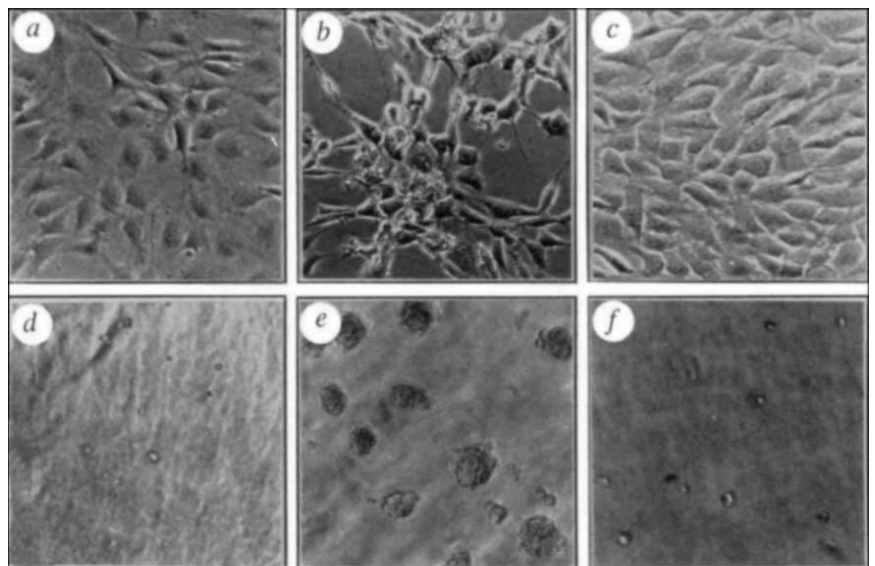
Cells transformed by ODC cDNA proliferated more rapidly than the untransfected controls, whereas cells carrying the ODC antisense construct had a markedly reduced growth rate (Fig. 2a). ODC cDNA-transfected cells were also able to grow in low serum (0.5% FCS) but antisense cDNA transfectants and control cells grew poorly under these conditions (Fig. 2b).

Expression of human ODC by the pHβAPr1neo vector (see Fig. 1 legend) transformed cells with a lower efficiency than the pLTRpoly/ODC construct, but several transformed foci were still obtained. Likewise, transfectants with ODC cDNA inserted into a dexamethasone-inducible pMAMneo vector caused morphological transformation in a fraction of cells after induction. Cells transfected with the pHβAPr1neo and pMAMneo plasmids carrying ODC cDNA in reverse orientation were all flat (data not shown).

Southern blot analysis of genomic DNA from the transfectants confirmed that sequences specific to human ODC were stably

FIG. 1 Morphology and soft agar growth of NIH3T3 cells transfected with pLTRpoly expression vector containing human ODC cDNA in sense and antisense orientations. Phase-contrast micrographs of parental NIH3T3 cells (a, d), NIH3T3 cells expressing human ODC (b, e) and antisense ODC mRNA (c, f) grown in monolayer cultures (a–c) and soft agar (d–f).

**METHODS.** Expression plasmid constructions: 1.8-kb *EcoRI* insert of human ODC cDNA was isolated from the pODC10/2H plasmid<sup>29</sup> and inserted in both orientations into the *EcoRI* site of the pLTRpoly vector<sup>30</sup> containing the U3 region of long terminal repeat (LTR) of Moloney murine leukaemia virus and the SV40 splicing and polyadenylation signals. We also subcloned the ODC cDNA into the expression vector pMAMneo (Clontech); vector contains the Rous sarcoma virus LTR enhancer linked to the dexamethasone-inducible mouse mammary tumour virus LTR promoter and SV40 splicing and polyadenylation sites, plus *neo*, the neomycin-resistance gene and into the pHβAPr1neo expression plasmid<sup>31</sup> (containing the human β-actin promoter plus intervening sequence-1 with enhancer-like activity, and the SV40 polyadenylation signal). To do this, the 1.8 kb *EcoRI* insert of the pODC10/2H plasmid was blunted with Klenow fragment, ligated to *SalI* linkers, and introduced into the *SalI* sites of the vector cloning boxes. Cell culture: NIH3T3 cells were grown to half-confluence in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal calf serum (FCS) and transfected with 5 μg of the pLTRpoly/ODC cDNA (sense and antisense) constructs and 250 ng pSV2neo using the TransfectACE reagent (Gibco, BRL) according to the instructions of the manufacturer. Transfected cells



were selected for resistance to G418 (500 μg ml<sup>-1</sup>) for two weeks. To determine their ability for anchorage-independent growth, 10<sup>4</sup> cells were suspended in 0.25% agar solution containing DMEM plus 10% FCS and overlaid onto 0.7% agar in 24-well (diameter 1.5 cm) plates (Nunc). The soft-agar colonies were photographed after 8 days of culture. The efficiency of colony formation of the ODC cDNA (sense) transfectants was 10–15% in four separate experiments.



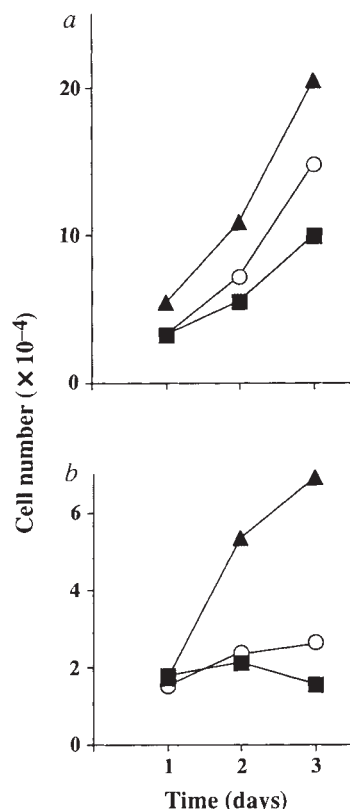


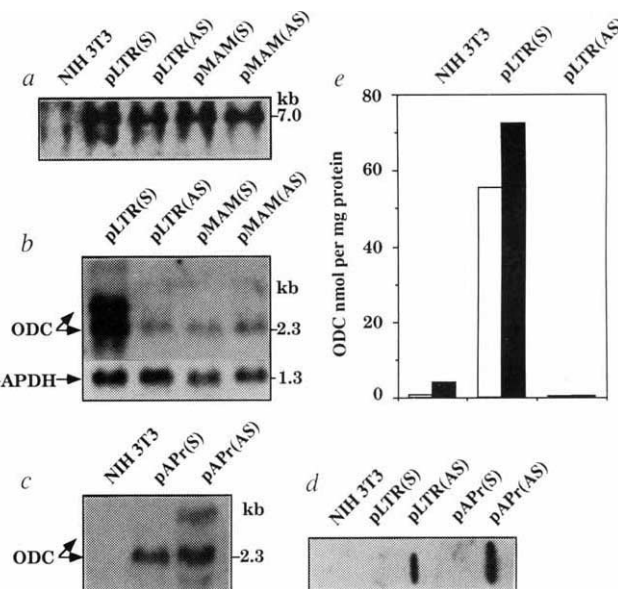
FIG. 3 Integration and expression of the human ODC cDNA in sense (S) and antisense (AS) orientations in NIH3T3 cell transfectants. **a**, Southern blot analysis of genomic DNA for human ODC sequences. Cells transfected with pLTRpoly/ODC and pMAMneo/ODC constructs showed one major *Bam*HI fragment of 7 kb (of roughly the same intensity) consisting of multiple copies of the ODC cDNA; pH $\beta$ Apr1neo/ODC transfectants gave a low copy number with different integration sites (not shown). **b–d**, Northern blot and slot-blot analyses of ODC mRNA. The size (2.3 kb) of the major endogenous mouse ODC mRNA is indicated. Transcript levels from pMAMneo/ODC transfectants represent the uninduced state and mainly reflect the endogenous ODC mRNA content in NIH3T3 cells (**b**). The predicted size of the chimaeric ODC mRNA resulting from the pH $\beta$ Apr1neo/ODC-transfected cells is indistinguishable from the endogenous mouse mRNA (**c**; 2.3 kb). The larger antisense ODC RNA transcribed under the human  $\beta$ -actin promoter might result from unsplicing of the intervening sequence (as a faint band of this size is also seen in the sense RNA-expressing cells on a longer exposure). In the antisense sequence of ODC cDNA there is a polyadenylation signal after 145 nucleotides, which predictedly results in  $\leq 0.2$  kb antisense RNAs that are difficult to detect by northern blotting (because they are not efficiently precipitated by ethanol and retained by the filter). **e**, ODC activity in nmol <sup>14</sup>CO<sub>2</sub> released per mg protein per hour. Open bars, serum-starved cells, filled bars, serum-stimulated cells.

**METHODS.** High molecular mass DNA (10  $\mu$ g) isolated from the different cell lines was digested with *Bam*HI, electrophoresed in 0.8% agarose gels, blotted onto nylon filters (Hybond-N, Amersham) and hybridized with a [<sup>32</sup>P]dCTP random-prime labelled human ODC insert (1.8-kb *Eco*RI fragment from the pODC10/H2 plasmid<sup>29</sup>) under stringent hybridization conditions. Genomic DNA was also digested with *Eco*RI, which released the 1.8-kb ODC cDNA insert from the vector of the pLTRpoly/ODC transfectants (not shown). For northern blot analysis (in **b** and **c**), total cellular RNA was extracted from the exponentially growing cells (after 2 days subculture) and 30- $\mu$ g aliquots were electrophoresed in a 1% agarose-formaldehyde gel, blotted and hybridized to the human ODC insert (and mouse pODC16 probe<sup>32</sup>; data not shown) and, as control, to a glyceraldehyde-3-phosphate dehydrogenase (GAPDH) probe<sup>11</sup>. To distinguish chimaeric ODC RNAs in the pH $\beta$ Apr1neo/ODC-transfected cells from endogenous ODC mRNA (both are 2.3 kb), the filter was washed using stringent conditions (0.1  $\times$  SSPE-0.1% SDS for 1.5 h at 60  $^{\circ}$ C; 1  $\times$  SSPE contains 10 mM sodium phosphate, pH 7.7, 0.18 M NaCl, 1 mM Na<sub>2</sub>EDTA) (**c**). As only faint signals of smaller-sized RNAs ( $\leq 0.2$  kb) were detected by northern blotting, we did a slot-blot analysis of RNA as described in ref. 33 with minor modifications (**d**). In brief, 10<sup>7</sup> cells were lysed in 200  $\mu$ l NP-40 buffer (10 mM Tris-HCl,

FIG. 2 Growth of the cells overexpressing ODC and antisense ODC mRNA in media with **a**, high and **b**, low serum concentrations.  $\circ$ , Parental NIH3T3 cells;  $\blacktriangle$ , cells expressing ODC;  $\blacksquare$ , cells expressing antisense ODC. **METHODS.** 3  $\times$  10<sup>4</sup> cells transfected with the pLTRpoly/ODC constructs were plated on 50-mm dishes in DMEM supplemented with 20% FCS (**a**) or 0.5% FCS (**b**). Cells were counted at the indicated times with a Coulter cell counter. All measurements were made in triplicate.

integrated (Fig. 3a, and data not shown). Northern blot analysis of total cellular RNA revealed that cells with the pLTRpoly/ODC construct express very high levels of human ODC mRNA (Fig. 3b). In addition to the endogenous mouse ODC mRNA (2.3 kilobases (kb)), two major chimaeric mRNA species of about 1.9 kb and 2.9 kb, which are the sizes predicted from the use of the polyadenylation signals in the human ODC cDNA itself or in SV40, respectively, and a minor larger RNA (possibly derived from a rearranged ODC cDNA) were detected. Cells with the pLTRpoly/ODC antisense construct did not significantly express the 2.9-kb RNA, but a short RNA of 0.2 kb was evident as a result of an early polyadenylation signal in the antisense sequence. This is read through in the pH $\beta$ Apr1neo construct (although the 0.2-kb RNA also appears to be formed), giving rise to an antisense RNA of 2.3 kb, which is the predicted full-length size (as it is for sense RNA), and minor differentially spliced or rearranged RNA species (Fig. 3c). The identity of antisense ODC transcripts was also confirmed by specific hybridization to a single-stranded oligonucleotide probe (Fig. 3d).

Cells carrying the pLTRpoly/ODC construct showed very



pH 7.4, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 0.5% NP-40), and nuclei were removed by centrifugation at 12,000g for 3 min at 4  $^{\circ}$ C. To the supernatant was added 120  $\mu$ l 20  $\times$  SSC (1  $\times$  SSC contains 0.15 M NaCl, 0.0125 M sodium citrate, pH 7) and 80  $\mu$ l 37% formaldehyde, the mixture was incubated for 15 min at 60  $^{\circ}$ C and the RNAs blotted onto a Hybond N filter (Amersham) using a Minifold II filtration manifold (Schleicher & Schuell). For specific detection of ODC antisense RNA, the filter was hybridized overnight to a synthetic 35-mer oligonucleotide (5'-GAACAAGCATTTGTAGCTTGACAA-TGGCAGAATG-3'; corresponding to nucleotides 1,722–1,756 in ODC cDNA) end-labelled with [ $\gamma$ -<sup>32</sup>P]ATP and T4 polynucleotide kinase. The filter was washed three times in 2  $\times$  SSC for 10 min and finally twice in 0.1  $\times$  SSC and 0.1% SDS for 15 min at room-temperature. The activity of ODC was measured from parallel cultures using [1-<sup>14</sup>C]-L-ornithine (50  $\mu$ M) as the substrate<sup>11</sup> (**e**).

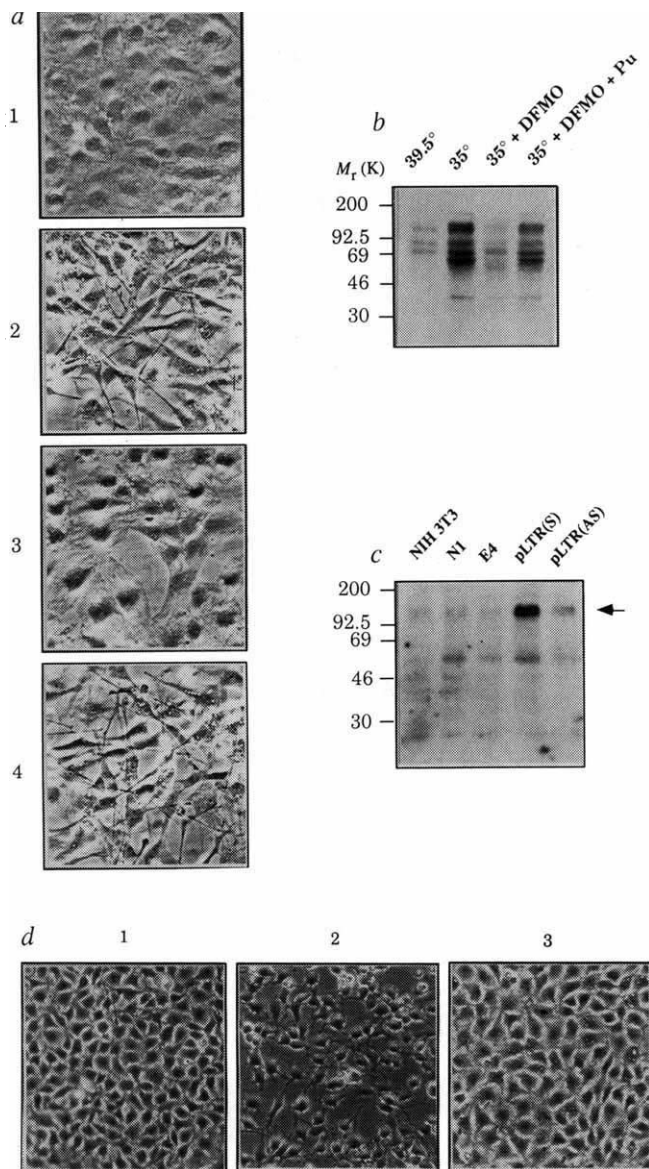


FIG. 4 Effect of inhibition of ODC induction on the morphology and protein tyrosine phosphorylation of temperature-sensitive (ts) RSVLA29 Rat-1 and rat 2R cells after temperature shift and analysis of tyrosine-phosphorylated proteins in NIH3T3 cells transformed by overexpression of human ODC. *a*, Phase contrast micrographs of Rat-1 tsRSVLA29 cells grown at 39.5 °C (1); shifted to 35 °C (2); shifted to 35 °C in the presence of DFMO (3); or transferred to 35 °C in the presence of DFMO and putrescine, the reaction product of ODC (4). *b*, Effect of DFMO and putrescine (Pu) on protein tyrosine phosphorylation in Rat-1 tsRSVLA29 cells in response to the temperature shift. Migration of *M<sub>r</sub>* standards in *b* and *c* is shown on the left. *c*, Pattern of tyrosine-phosphorylated proteins in parental NIH3T3 cells, and NIH3T3 cells transfected with pSV2neo alone (N1) or together with human ODC cDNA in sense (pLTR(S)) or antisense (pLTR(AS)) orientations, or with the *c-Ha-ras*<sup>Val12</sup> oncogene (E4 cells<sup>11</sup>). Arrow indicates 130K band (see text). *d*, Micrographs of 2R cells growing at the restrictive temperature (1), or shifted to the permissive temperature (2); and of 2R cell transfectants expressing antisense ODC RNA downshifted to the permissive temperature (3).

**METHODS.** *a*, Rat-1 tsRSVLA29 cells were grown in DMEM containing 5% newborn calf serum (NBCS) at the restrictive temperature (39.5 °C) for 3 days. Cells were then fed with fresh DMEM plus 5% dialysed NBCS, 2 mM DFMO and 100 µM putrescine were added to the cultures indicated, and the appropriate cultures transferred to a 35 °C incubator and photographed after 16 h. *b*, The cells were lysed at 12 h, equal amounts of proteins resolved by SDS-PAGE, transferred to nitrocellulose and immunoblotted with affinity-purified antiphosphotyrosine antibodies as described<sup>34</sup>, using <sup>125</sup>I-labelled protein-A (Amersham; 1 mCi ml<sup>-1</sup>) for detection. Filters were exposed to Kodak XAR-2 film with an intensifying screen at -70 °C for 20 h. *c*, Parental NIH3T3 cells and the NIH3T3 transfectants were grown for 2 days in DMEM plus 10% FCS and analysed for tyrosine-phosphorylated proteins. *d*, 2R cells (carrying the B77 *src*-gene mutant LA339) and 2R cells transfected with the pHβApr1neo/ODC cDNA antisense construct were grown in DMEM containing 5% NBCS at the restrictive temperature (39.5 °C) for 2 days, then cultures were cooled to 35 °C and photographed after 24 h. A similar result was obtained with 2R cells transfected with pLTRpoly/antisense ODC cDNA.

high constitutive ODC activity (Fig. 3e) and the pHβApr1neo and pMAMneo constructs yielded about 60 and 30% of that activity, respectively. In all the (ODC cDNA sense) transfectants, ODC activity remained high even after serum starvation, whereas in normal NIH3T3 cells it was reduced to barely detectable levels (Fig. 3e, and data not shown).

By which mechanism(s) does ODC induce cell transformation? We have recently found that inhibition of ODC in Rous sarcoma virus (RSV)-infected rat 2R cells by  $\alpha$ -difluoromethyl ornithine (DFMO), a specific inhibitor of ODC<sup>3,18</sup>, prevents both the disintegration of actin filaments and morphological transformation (E.H. and L.C.A., manuscript submitted). Further studies of this and another temperature-sensitive (ts) RSV mutant (RSVLA29 Rat-1) cell line verified that blocking of ODC by DFMO or by transcription of ODC antisense RNA prevents transformation (Fig. 4a, d). Because the *v-src* transforming gene of RSV encodes a tyrosine-specific kinase<sup>19,20</sup>, and as we found that DFMO causes a reduction in protein tyrosine phosphorylation in LA29 cells (Fig. 4b), we investigated whether the increase in production of ODC and in its transforming activity was associated with changes in tyrosine phosphorylation. Immunoblotting of total cellular protein with antiphosphotyrosine antibodies revealed one band of about 130K which had undergone markedly increased tyrosine phosphorylation in the ODC trans-

fectants compared with the parental NIH3T3 cells or cells transformed with *c-Ha-ras*<sup>Val12</sup> (Fig. 4c). Attempts to identify this protein as Ras-GAP (GTPase-activating protein)<sup>21</sup> have been unsuccessful.

Our findings that blocking of ODC inhibits transformation of rat fibroblasts by *v-src* and that reciprocally constitutive overexpression of human ODC induces morphological transformation of NIH3T3 cells and Rat-1 fibroblasts (results not shown), indicate that ODC is both necessary and sufficient for transformation. The use of different expression vectors for ODC shows further that the degree of transformation correlates with the expression of ODC after it has reached a certain threshold. This might explain a report<sup>22</sup> in which cell transformation was only successful when ODC was co-expressed with the *c-Ha-ras* oncogene and not with overexpressed murine ODC alone.

Deregulated expression of ODC and polyamine metabolism has been observed in a variety of animal and human malignancies<sup>23-27</sup>. Our results indicate that the aberrant expression of ODC is not just a coincident, pleiotypic response to transformation, but a critical factor contributing to oncogenesis. The ODC gene should therefore be recognized as a member of the growing family of cellular proto-oncogenes. Positioning of ODC at the convergence point of the signalling pathway of many oncogenes suggests that ODC might be transducing their transforming



activity which would make it a possible target for therapeutic interventions<sup>28</sup>. □

Received 17 June; accepted 9 October 1992.

1. Tabor, C. W. & Tabor, H. A. *Rev. Biochem.* **53**, 749–790 (1984).
2. Pegg, A. E. *Biochem. J.* **234**, 249–262 (1986).
3. Heby, O. & Persson, L. *Trends biochem. Sci.* **15**, 153–158 (1990).
4. Steglich, C., Grens, A. & Scheffler, I. E. *Somatic Cell molec. Gen.* **11**, 11–23 (1985).
5. Pohjanpelto, P., Hölttä, E. & Jänne, O. A. *Molec. cell Biol.* **5**, 1385–1390 (1985).
6. Yuspa, S. H. et al. *Nature* **262**, 402–404 (1976).
7. Gilmour, S. K., Verma, A. K., Madara, T. & O'Brien, T. G. *Cancer Res.* **47**, 1221–1225 (1987).
8. Don, S. & Bachrach, U. *Cancer Res.* **35**, 3618–3622 (1975).
9. Gazdar, A. F., Stull, H. B., Kilton, L. J. & Bachrach, U. *Nature* **262**, 696–698 (1976).
10. Haddox, M. K., Magun, B. E. & Russell, D. H. *Cancer Res.* **40**, 604–608 (1980).
11. Hölttä, E., Sistonen, L. & Alitalo, K. *J. biol. Chem.* **263**, 4500–4507 (1988).
12. Chang, B. K., Libby, P. R., Bergeron, R. & Porter, C. W. *Biochem. biophys. Res. Commun.* **157**, 264–270 (1988).
13. Sistonen, L., Hölttä, E., Mäkelä, T. P., Keski-Oja, J. & Alitalo, K. *EMBO J.* **8**, 815–822 (1989).
14. Sistonen, L., Hölttä, E., Lehtola, H., Lehtola, L. & Alitalo, K. *J. Cell Biol.* **109**, 1911–1919 (1989).
15. Pohjanpelto, P., Virtanen, I. & Hölttä, E. *Nature* **293**, 475–477 (1981).
16. Balasundaram, D., Tabor, C. W. & Tabor, H. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5872–5876 (1991).
17. Southern, P. J. & Berg, P. *J. molec. appl. Genet.* **1**, 327–341 (1982).
18. Metcalf, B. W. et al. *J. Am. chem. Soc.* **100**, 2551–2553 (1978).
19. Erikson, R. L., Purchio, A. F., Erikson, E., Collett, M. S. & Bruggs, J. S. *Cell Biol.* **87**, 319–325 (1980).
20. Sefton, B. M., Hunter, T., Beemon, K. & Eckhart, W. *Cell* **20**, 807–816 (1980).
21. Ellis, C., Moran, M., McCormick, F. & Pawson, T. *Nature* **343**, 377–381 (1990).
22. Hibshoosh, H., Johnson, M. & Weinstein, I. B. *Oncogene* **6**, 739–743 (1991).
23. Jänne, J., Hölttä, E., Kallio, A. & Käpyaho, K. *Spec. Top. endocrin. Metab.* **5**, 227–293 (1983).
24. Luk, G. D. & Baylin, S. B. *New Engl. J. Med.* **311**, 80–83 (1984).
25. Pegg, A. E. *Cancer Res.* **48**, 759–774 (1988).
26. Katz, A. & Kahana, C. *EMBO J.* **8**, 1163–1167 (1989).
27. Tonin, P. N., Yeger, H., Stallings, R. L., Srinivasan, P. R. & Lewis, W. H. *Oncogene* **4**, 1117–1121 (1989).
28. Seiler, N. et al. *Cancer Res.* **50**, 5077–5083 (1990).
29. Hickock, N. J., Seppänen, P. J., Gunsalus, G. L. & Jänne, O. A. *DNA* **6**, 179–187 (1987).
30. Mäkelä, T. P., Partanen, J., Schwab, M. & Alitalo, K. *Gene* **118**, 293–294 (1992).
31. Gunning, P., Leavitt, J., Muscat, G., Ng, S.-Y. & Kedes, L. *Proc. natn. Acad. Sci. U.S.A.* **84**, 4831–4835 (1987).
32. Jänne, O. A., Kontula, K. K., Isomaa, V. V. & Bardin, C. W. *Ann. N.Y. Acad. Sci.* **438**, 72–84 (1984).
33. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1989).
34. Kamps, M. P. & Sefton, B. M. *Oncogene* **2**, 305–315 (1988).

**ACKNOWLEDGEMENTS.** We thank O. A. Jänne for the pODC10/2H plasmid, T. P. Mäkelä for the pLTRpoly vector, J. A. Wyke for the tsRSVLA29 Rat-1 and 2R cells, B. Sefton for the antiphosphotyrosine antibodies, and A. M. Siivonen and A. Asikainen for technical assistance. DL- $\alpha$ -difluoromethyl ornithine was a gift from M. Merrell Dow. This work was supported by the Finnish Academy of Sciences, University of Helsinki, the Sigrid Jusélius Foundation and the Finnish Cancer Organization.

## Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160

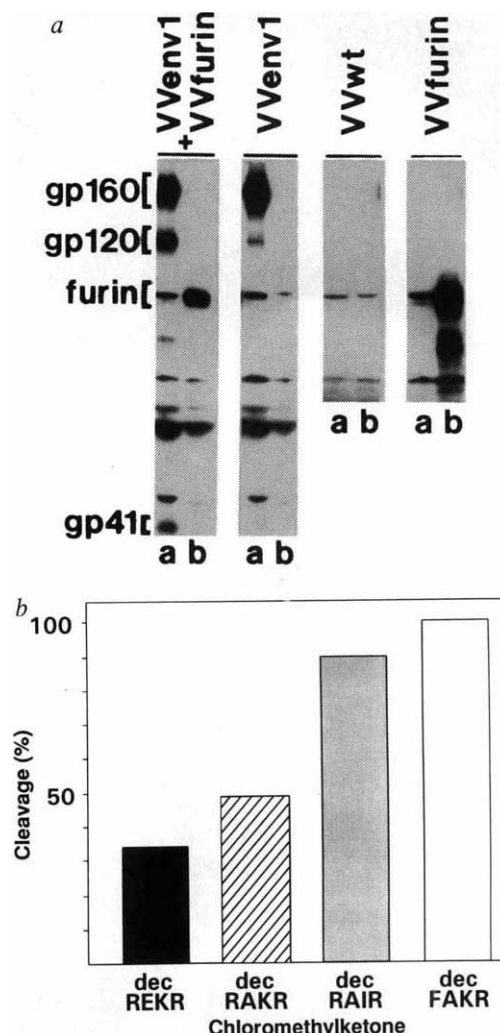
Sabine Hallenberger\*, Valerie Bosch†, Herbert Angliker‡, Elliott Shaw‡, Hans-Dieter Klenk\* & Wolfgang Garten\*

\* Institut für Virologie, Philips-Universität Marburg, Robert-Kochstrasse 17, 3550 Marburg, Germany

† Institut für Angewandte Tumorstudiologie am Deutschen Krebsforschungszentrum Heidelberg, Im Neuenheimer Feld 242, 6900 Heidelberg 1, Germany

‡ Friedrich Miescher-Institut Basel, PO Box 2543, CH-4002 Basel, Switzerland

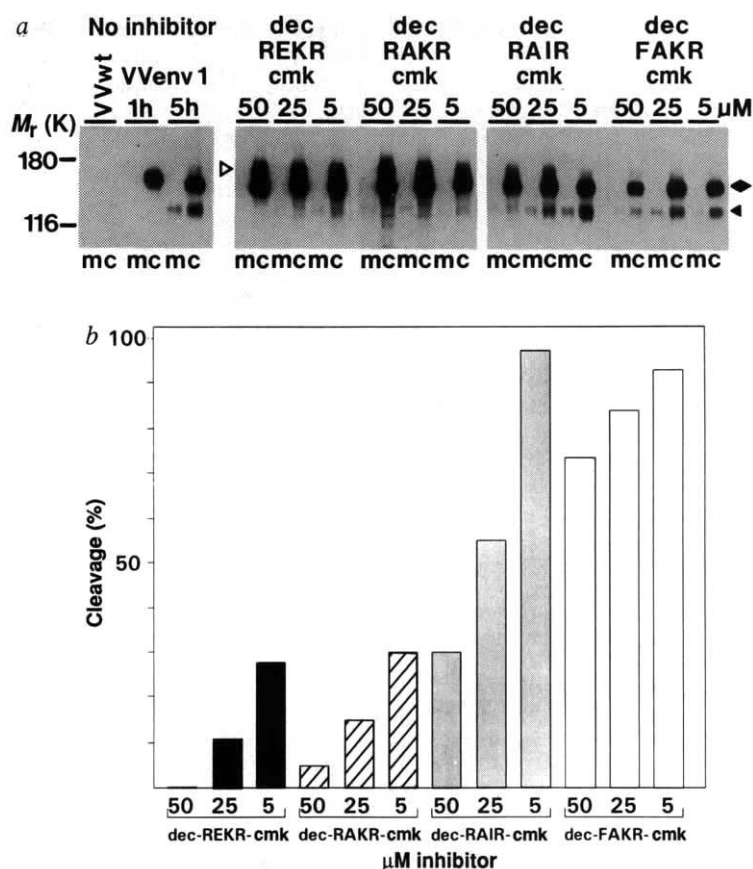
**THE envelope glycoprotein of human immunodeficiency virus (HIV) initiates infection<sup>1</sup> by mediating fusion of the viral envelope with the cell membrane. Fusion activity requires proteolytic cleavage of the gp160 protein into gp120 and gp41 at a site containing several arginine and lysine residues<sup>2</sup>. Activation at basic cleavage sites is observed with many membrane proteins of cellular and viral origin. We have recently found that the enzyme activating the haemagglutinin of fowl plague virus (FPV), an avian influenza virus, is furin<sup>3</sup>. Furin, a subtilisin-like eukaryotic endoprotease<sup>4–6</sup>, has a substrate specificity for the consensus amino-acid sequence Arg-X-Lys/Arg-Arg at the cleavage site<sup>7</sup>. We show here that the glycoprotein of HIV-1, which has the same protease recognition motif as the FPV haemagglutinin, is also activated by furin.**



**FIG. 1** Cleavage of the HIV-1 glycoprotein by human furin expressed from a vaccinia virus vector and its inhibition by sequence-specific peptidyl-chloromethylketones. **a**, Coexpression of furin and gp160 by recombinant vaccinia viruses. Proteins immunoprecipitated with anti-HIV serum (a) or anti-furin serum (b) from lysates of CV-1 cells are shown. Cells were doubly infected with recombinant vaccinia viruses VVenv1<sup>8</sup> and VVfurin<sup>5</sup> or singly infected with each recombinant and wild-type vaccinia virus (VVwt). **b**, Cleavage inhibition by peptidyl chloromethylketones (cmk). The effect of the inhibitors decanoyl-arginyl-glutamyl-lysyl-arginyl-chloromethylketone (decREKR), decanoyl-arginyl-alanyl-lysyl-arginyl-chloromethylketone (decRAKR), decanoyl-arginyl-alanyl-isoleucyl-arginyl-chloromethylketone (decRAIR), and decanoyl-phenylalanyl-alanyl-lysyl-arginyl-chloromethylketone (decFAKR; H.A., R. Wilkstrom, E.S., W. Bannan and R. S. Fuller, manuscript submitted) were analysed at a concentration of 50  $\mu$ M in cells coinfecting with VVenv1 and VVfurin.

**METHODS.** Multiplicities of infection (m.o.i.) were 5 PFU per cell. At 19 h post infection (p.i.), cells were starved of methionine and cysteine for 1 h and then labelled with [<sup>35</sup>S]methionine and [<sup>35</sup>S]cysteine for 2 h. After a 3-h chase period, cells were lysed in 1 ml lysis buffer<sup>3</sup>. Equal volumes of the cellular lysates were exposed for 1 h at 4° to either pooled sera of healthy HIV-1 seropositive donors (anti-HIV antiserum<sup>15</sup>; NIH AIDS Research and Reference Reagent Program, Rockville, MD) or a rabbit antiserum against furin. Antisera were used undiluted in all experiments to accomplish quantitative immunoprecipitation. Protein A-Sepharose beads (Sigma) were then added for 2 h at 4 °C. Precipitates were analysed by electrophoresis on 8% polyacrylamide gels in the presence of SDS followed by fluorography. Cleavage inhibitors were added 1 h and replenished at 10 h p.i.<sup>16</sup>. Control samples were incubated without inhibitors. After SDS-PAGE and fluorography, the extent of cleavage was determined as follows: radioactive bands were cut out of the gel, rehydrated in water and then solubilized in Soluene-350 (Packard) for 1 h at 60 °C and counted in liquid scintillator (Zinsser Analytic). The extent of cleavage was determined as the amount of gp120 observed in the presence of inhibitors relative to the amount of gp120 observed in the absence of these inhibitors.

FIG. 2 Effects of peptidyl-chloromethylketones on gp160 cleavage mediated by the endogenous protease of HeLa cells. **a**, Inhibition of cleavage. HIV glycoprotein immunoprecipitated from medium (m) and cell lysates (c) was analysed. Cells incubated without inhibitors were either lysed immediately after the pulse (1 h) or after a 5-h chase period. Cells incubated with inhibitors were analysed only after pulse-chase labelling. Open arrowhead indicates gp160 with mature carbohydrate side chain;  $\blacklozenge$ , gp160 with immature carbohydrate side chains; filled arrowheads, gp120. HeLa T4 cells were infected with either VVwt or VVenv1 at m.o.i. of 10. **b**, Quantification of cleavage inhibition. The extent of cleavage in the samples shown in **a** is measured as the amount of gp120 observed in the presence of inhibitors relative to the amount of gp120 observed in the absence of inhibitors. **METHODS**. After a 1-h adsorption period, DMEM containing various concentrations of the different inhibitors was added. At 10 h p.i., cells were starved of methionine and cysteine for 1 h and subsequently exposed to pulse-chase labelling with [ $^{35}$ S]methionine and [ $^{35}$ S]cysteine. HIV glycoprotein was immunoprecipitated with human anti-HIV antiserum from media and cell lysates and analysed by SDS-PAGE.



Furthermore, we present evidence that peptidyl-chloromethyl ketones that have the Arg-X-Lys/Arg-Arg motif and which are specific inhibitors of furin<sup>3</sup> interfere with cleavage of the HIV glycoprotein and hence its activation and the formation of infectious virus particles.

To determine whether cleavage of the HIV-1 precursor glycoprotein gp160 is mediated by furin, gp160 complementary DNA<sup>8</sup> and human furin cDNA isolated from a hepatoma (HepG2) cDNA library<sup>5</sup> were expressed in CV-1 cells alone or in combination using vaccinia virus recombinants as vectors (Fig. 1). Previous studies on the expression of nerve growth factor  $\beta$ -NGF<sup>5</sup> and on FPV haemagglutinin<sup>3</sup> had shown that late after infection with the vaccinia vectors, cleavage by the endogenous CV-1 protease ceased and the expression products accumulated in the uncleaved form. Similarly, only gp160 was observed when cells were labelled at 19h after infection with VVenv1 alone, but after coexpression with human furin, which is a protein of  $M_r$ 92,000 (92K)<sup>3</sup>, the cleavage products gp120 and gp41 were detected. Therefore furin is able to process the HIV-1 glycoprotein. The finding that only ~40% of gp160 is cleaved can be explained by the observation that a relatively high proportion of the glycoprotein is retained in the endoplasmic reticulum, presumably as a result of incorrect folding<sup>9</sup>. Thus, only a minority of gp160 molecules appear to be able to reach the Golgi complex of the *trans* Golgi compartment, where gp160 cleavage is supposed to occur<sup>10-12</sup> and where furin has been localized<sup>5</sup>. Figure 1b shows that chloroalkylketones with an appropriate peptide moiety<sup>3</sup> (decanoyl-REKR-cmk and decanoyl-RAKR-cmk; one-letter amino-acid code) interfere with the cleavage of gp160 by overexpressed furin, whereas compounds not having the correct consensus sequence (decanoyl-RAIR-cmk and decanoyl-FAKR-cmk) are inactive. Therefore specific inhibitors of gp160 cleavage are now available.

To test whether peptidyl-chloromethylketones are able to inhibit endogenous furin present in human cells, their effect on

HIV-1 glycoprotein expressed in HeLa-CD4 cells by a vaccinia virus vector was analysed (Fig. 2a, b). When pulse-chase labelled at 10 h post-infection cleavage of gp160 and release of gp120 into the culture medium can be clearly detected. In the presence of 5  $\mu$ M decanoyl-REKR-cmk and decanoyl-RAKR-cmk, cleavage is inhibited and the precursor that is not released into the medium has acquired a higher electrophoretic mobility. As the precursor is resistant to endoglucosaminidase H under these conditions, the higher molecular weight probably reflects

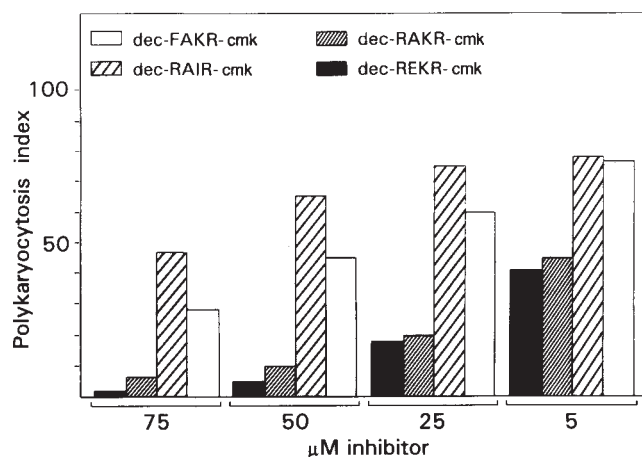


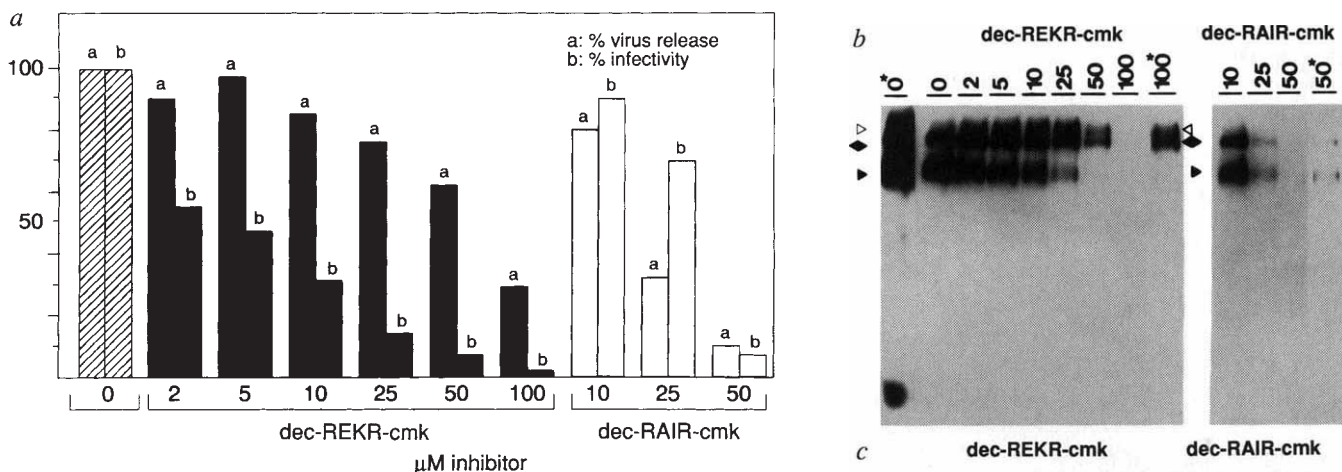
FIG. 3 Effect of peptidyl-chloromethylketones on HIV glycoprotein-induced syncytia formation in HeLa T4 cells. Confluent monolayers of HeLa T4 cells were infected with either VVwt or VVenv1 at m.o.i. of 2.5. After 60-min adsorption, viral inoculum was removed and medium containing inhibitors was added. At 8 h p.i. cell cultures were examined by phase-contrast microscopy (magnification, 250 $\times$ ). The polykaryocytosis indices (PI) (percentage of total nuclei of a monolayer present in polykaryons) are indicated by columns. PI of VVenv1 infected cells without inhibitor was  $\geq 80$ ; PI of VVwt infected cells was  $< 5$ .



maturation of the oligosaccharides to more elongated forms. Decanoyl-RAIR-cmk and decanoyl-FAKR-cmk do not show these effects. These findings support the concept that cleavage of gp160 is specifically blocked by furin inhibitors, and they suggest that the inhibitors do not interfere with other processing events (see below).

To elucidate the effects of cleavage inhibition on biological functions of the envelope glycoprotein, we first analysed syncytia formation in HeLa-CD4 cells. Figure 3 shows the inhibitory potential of the compounds. Again there was a distinct effect with decanoyl-RAKR-cmk and decanoyl-REKR-cmk, which completely blocked cell fusion at  $\geq 50 \mu\text{M}$ . The other compounds barely inhibited fusion. Decanoyl-FAKR-cmk, as compared to decanoyl-RAIR-cmk, had retained some inhibitory activity which, however, seems to reflect the relative toxicity of this compound (compare Fig. 2a) and is therefore probably not specific. To test whether the inhibitors also interfered with virus infectivity, we analysed the effects of decanoyl-REKR-cmk and decanoyl-RAIR-cmk on the replication of HIV-1 in human

lymphocytes. With increasing doses of decanoyl-REKR-cmk, virion release decreases only slightly, whereas infectivity drops sharply. Thus, at  $50 \mu\text{M}$  decanoyl-REKR-cmk, the infectivity of the released virus is about 10-fold lower than that of untreated control virus. In contrast, in the presence of decanoyl-RAIR-cmk, there is a drastic reduction in released particles, but their infectivity is not reduced. These observations again underline the potential of decanoyl-REKR-cmk as a specific inhibitor of proteolytic activation, whereas decanoyl-RAIR-cmk shows not only less specificity but also considerable toxicity. Viral glycoproteins present in MT-4 cells (Fig. 4b) and virus particles (Fig. 4c) were analysed after inhibitor treatment. In agreement with the data shown in Fig. 2a, there is an increase in the gp160:gp120 ratio in the cell lysates with increasing decanoyl-REKR-cmk concentrations. The increased molecular weight of gp160, which presumably reflects carbohydrate maturation (Fig. 2a), can also be seen. With the dibasic compound decanoyl-RAIR-cmk, the gp160:gp120 ratio remains virtually constant and the larger form of gp160 does not appear. This again



**FIG. 4** Effects of cleavage inhibitors on formation of infectious virus. Formation of HIV-1 particles (strain HTLV-IIIB; ref. 17) was examined in the transformed T-lymphocyte cell line MT-4 (ref. 18). **a**, Infectivity of virus released from inhibitor-treated MT-4 cells. Using an ELISA test, it was established that in the inhibitor-treated cultures 50–80% of the total antigen in the medium was precipitable by polyethylene glycol and so are virus particles<sup>19</sup>. This indicates that there has been no significant release of soluble antigen from the cultures as a result of cell damage, so we interpreted the decrease in HIV-1 released into the medium at higher inhibitor concentrations as a measure of the general cytotoxic effect of such inhibitors on MT-4 cells. The infectivity of virus particles released from inhibitor-treated cultures was analysed by infecting fresh MT-4 cells with equal amounts of virus as determined by ELISA. **b**, Viral glycoproteins in inhibitor-treated MT-4 cells and **c**, in released virus particles. The positions of the viral glycoproteins gp160 ( $\blacklozenge$ ), gp120 ( $\blacktriangle$ ), gp41 ( $\blacksquare$ ), and gp160 with mature oligosaccharide side chains ( $\Delta$ ) are indicated. An immunoprecipitated component observed in the virus, migrating at about 75K, presumably represents a proteolytic cleavage product of gp120. 0\*, 100\* (decREKR-cmk) and 50\* (decRAIR-cmk) indicate longer exposures of lanes 0, 100 (decREKR-cmk) and 50 (decRAIR-cmk), respectively. The amount of labelled glycoprotein decreases with increasing inhibitor concentration. A part of this decrease can be accounted for by the decrease in the amount of released HIV-1 with increasing inhibitor concentration, see **a**. But at the highest concentrations of both inhibitors, the reduction in the amount of labelled glycoprotein is higher than the reduction of released virus. This may indicate a differentially greater effect on the viral glycoproteins.

**METHODS.** Rapid synchronous replication was achieved by adding 5% of a completely infected MT-4 culture (at 3 days p.i.) to 95% fresh subconfluent MT-4 cells, cultivated in RPMI 1640 medium. At 20 h p.i. the infected cells were plated in 1-ml aliquots and different concentrations of inhibitors added; 6 h later cells were collected by low-speed centrifugation, the medium containing virus that had been synthesized before inhibitor treatment was discarded, and the cells were resuspended in 1 ml medium containing [ $^3\text{H}$ ]glucosamine ( $100 \mu\text{Ci ml}^{-1}$ ; Amersham–Buchler). Inhibitors were replen-

ished 7 h later. After a further 8 h, cells and media were collected and prepared for further analysis. Immunofluorescence using human AIDS serum revealed that all cells were infected. Infectivity assay of released virus: virus particles from  $800 \mu\text{l}$  medium were precipitated with polyethylene glycol (PEG) as described<sup>20</sup>. The amount of HIV antigen in the redissolved PEG pellets and in the original culture medium without PEG were determined using a standard p24-capture ELISA (Organon Teknica)<sup>21</sup>. Equal amounts of virus particles, as determined by ELISA of the PEG pellets from media of inhibitor-treated cultures, were used to infect fresh subconfluent MT-4 cells. At 5 h p.i., infected cultures were washed with medium to remove input virus and cultured further. At 31 h p.i., the amounts of newly synthesized virus released into the medium were determined again by ELISA. These values reflect the infectivity of the virus used for infection. The amount and infectivity of virus released from untreated cultures were each taken as 100%. Analysis of envelope glycoprotein: labelled cells and PEG-precipitated virus, both from  $800 \mu\text{l}$  culture treated with inhibitors at micromolar concentrations, were dissolved in lysis buffer (1% Triton-X100, 0.5% desoxycholate, 0.1% SDS, 2mM EDTA in 10 mM sodium phosphate, pH 7.2) and immunoprecipitated with rabbit antiserum specific for gp120 and gp160, followed by polyacrylamide gel electrophoresis<sup>22</sup>.

indicates that, unlike decanoyl-REKR-cmk, decanoyl-RAIR-cmk does not specifically affect proteolytic cleavage of gp160. Virus particles released from decanoyl-REKR-cmk-treated cells (Fig. 4c) contain increasing amounts of gp160 in its high-*M<sub>r</sub>* form. Whether or not gp160 is incorporated into virions has been the subject of some controversy<sup>13,14</sup>. The data presented here demonstrate that cleavage is not a prerequisite for the transport of the glycoprotein to the cell surface and its incorporation into virions. On the other hand, the reduced infectivity of virus produced in the presence of decanoyl-REKR-cmk is clearly a consequence of the inability of gp160 to induce fusion. The virus produced under these conditions still contained some gp120, indicating that cleavage inhibition was not complete. As cleavage is a late step in processing, it is not surprising that the proportion of cleaved glycoprotein was higher in virions than in cells. The coexistence of gp120 and gp160 in virus particles raises the possibility that cleaved and hence potentially functional glycoprotein may be negatively influenced by an association with gp160 to hetero-oligomers.

As decanoyl-REKR-cmk and decanoyl-RAIR-cmk are specific inhibitors of furin<sup>3</sup>, their interference with cleavage of gp160 supports the results obtained after expressing the enzyme from cDNA in identifying it as the activating protease. Our data show also that the inhibitors reduce the infectivity of the virus and may therefore have the potential to arrest the spread of infection in the organism. Furthermore, the inhibitors prevent shedding of gp120 from the surface of infected cells, which is also suspected to play a role in viral pathogenesis. Further

investigation will establish whether peptidyl-chloromethylketones that mimic the cleavage site of gp160 can be effective as therapeutic agents against HIV infection. □

Received 22 June; accepted 2 October 1992.

1. Dalglish, A. G. *et al.* *Nature* **312**, 763–767 (1984).
2. McCune, J. M. *et al.* *Cell* **53**, 55–67 (1988).
3. Stieneke-Grober, A. *et al.* *EMBO J.* **11**, 2407–2412 (1992).
4. von den Ouweland, A. M. W., van Duijnhoven, H. L. P., Keizer, G. D., Dorssers, L. C. J. & van de Ven, W. J. M. *Nucleic Acid Res.* **18**, 664 (1990).
5. Bresnahan, P. A. *et al.* *J. Cell Biol.* **111**, 2851–2859 (1990).
6. Wise, R. J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **87**, 9368–9382 (1990).
7. Hosaka, M. *et al.* *J. Biol. Chem.* **266**, 12127–12130 (1991).
8. Owens, R. J. & Compans, R. W. *J. Virol.* **63**, 978–982 (1989).
9. Willey, R. L., Bonifacio, J. S., Potts, B. J., Martin, M. A. & Klausner, R. D. *Proc. natn. Acad. Sci. U.S.A.* **85**, 9580–9584 (1988).
10. Dewar, R. L., Vasudevarani, M. B., Natarajan, V. & Sazman, N. P. *J. Virol.* **63**, 2452–2456 (1989).
11. Stein, B. S. & Engleman, E. G. *J. Biol. Chem.* **265**, 2640–2649 (1990).
12. Earl, P. L., Doms, R. W. & Moss, B. *Proc. natn. Acad. Sci. U.S.A.* **87**, 648–651 (1990).
13. Guo, H.-G. *et al.* *Virology* **174**, 217–224 (1990).
14. Willey, R. L., Klimkait, T., Frucht, D. M., Bonifacio, J. S. & Martin, M. A. *Virology* **184**, 319–329 (1991).
15. Prince, A. M. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **85**, 6944–6948 (1988).
16. Garten, W., Stieneke, A., Shaw, E., Wikstrom, P. & Klenk, H.-D. *Virology* **172**, 25–31 (1989).
17. Myers, G. *et al.* *Human retroviruses and AIDS* (Los Alamos National Laboratory, NM, 1991).
18. Harada, S., Koyanagi, Y. & Yamamoto, N. *Science* **229**, 563–566 (1985).
19. Mergener, K. *et al.* *Virology* **186**, 25–39 (1992).
20. Wilk, T., Mierswa, H., Kräusslich, H.-G., Dunn, J. & Bosch, V. *Virus Genes* **6**, 229–246 (1992).
21. Wilk, T., Pfeiffer, T. & Bosch, V. *Virology* **189**, 167–177 (1992).
22. Bosch, V. & Pawlita, M. *J. Virol.* **64**, 2337–2344 (1990).

ACKNOWLEDGEMENTS. We thank R. W. Compans and G. Thomas for vaccinia recombinants of the HIV-1 envelope protein and of human furin, respectively; W. Schäfer for preparing the furin-specific antiserum; S. Tucker, Birmingham, for discussion. Human HIV-1 immune globulin was obtained from A. Prince, Bethesda through the AIDS Research and Reference Reagent Program, NIH. This work was supported by grants from the BMFT-Forschungsförderung/Förderschwerpunkt AIDS to W.G. and V.B.

## Natural inhibitor of transforming growth factor- $\beta$ protects against scarring in experimental kidney disease

Wayne A. Border\*, Nancy A. Noble\*, Tatsuo Yamamoto\*, John R. Harper†, Yu Yamaguchi‡, Michael D. Pierschbacher‡‡ & Erkki Ruoslahti‡‡

\* Division of Nephrology, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA

† Telios Pharmaceuticals Inc., San Diego, California 92121, USA

‡ Cancer Research Center, La Jolla Cancer Research Foundation, La Jolla, California 92037, USA

THE central pathological feature of human kidney disease that leads to kidney failure is the accumulation of extracellular matrix in glomeruli. Overexpression of transforming growth factor- $\beta$  (TGF- $\beta$ ) underlies the accumulation of pathological matrix in experimental glomerulonephritis<sup>1</sup>. Administration of an antibody raised against TGF- $\beta$  to glomerulonephritic rats suppresses glomerular matrix production and prevents matrix accumulation in the injured glomeruli<sup>2</sup>. One of the matrix components induced by TGF- $\beta$ , the proteoglycan decorin, can bind TGF- $\beta$  and neutralize its biological activity<sup>3</sup>, so decorin may be a natural regulator of TGF- $\beta$  (refs 3, 4). We tested whether decorin could antagonize the action of TGF- $\beta$  *in vivo* using the experimental glomerulonephritis model<sup>1</sup>. We report here that administration of decorin inhibits the increased production of extracellular matrix and attenuates manifestations of disease, confirming our hypothesis. On the basis of our results, decorin may eventually prove to be clinically useful in diseases associated with overproduction of TGF- $\beta$ .

Three isoforms of TGF- $\beta$  are expressed in mammals, TGF- $\beta$ 1, 2 and 3 (ref. 5). As shown in Fig. 1, recombinant human decorin blocks the growth-inhibiting action of each TGF- $\beta$  isoform in

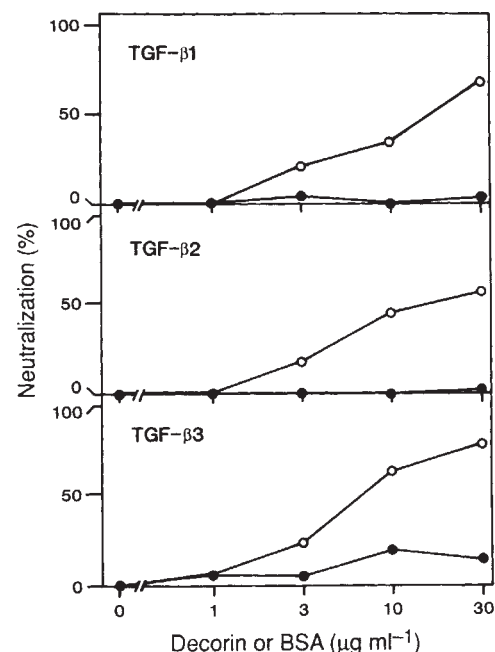


FIG. 1 Neutralization of the activity of TGF- $\beta$ 1, 2 and 3 by recombinant human decorin; TGF- $\beta$  inhibits mink lung cell growth. [<sup>3</sup>H]Thymidine incorporation was determined with TGF- $\beta$  alone (not shown) and with increasing concentrations of decorin (circles) or bovine serum albumin (BSA) (filled circles) as described<sup>3</sup>. Incorporation without TGF- $\beta$  or decorin is defined as 100% (125,000 c.p.m. for TGF- $\beta$ 1 and TGF- $\beta$ 3; 107,000 c.p.m. for TGF- $\beta$ 2); incorporation with TGF- $\beta$  alone is defined as 0% (57,000 c.p.m. for TGF- $\beta$ 1; 78,000 c.p.m. for TGF- $\beta$ 2; 59,000 c.p.m. for TGF- $\beta$ 3). Points represent means of duplicate samples.

METHODS. Purified human TGF- $\beta$ 1 and 2 were from R & D Systems (Minneapolis, MN); TGF- $\beta$ 3 was a gift from A. Roberts and M. Sporn. TGF- $\beta$  was added at concentrations that inhibited [<sup>3</sup>H]thymidine incorporation by 50%; 0.2 ng ml<sup>-1</sup>, 0.1 ng ml<sup>-1</sup> and 0.05 ng ml<sup>-1</sup> for TGF- $\beta$ 1, TGF- $\beta$ 2 and TGF- $\beta$ 3, respectively. Recombinant decorin and BSA were prepared as described for Fig. 2.



a mink lung epithelial cell assay. The anti-TGF- $\beta$ 1 serum used previously to ameliorate glomerulonephritis<sup>2</sup> neutralizes only TGF- $\beta$ 1 (results not shown). Thus decorin has a broader binding reactivity against TGF- $\beta$ s than does anti-TGF- $\beta$ 1.

Recombinant human decorin and decorin purified from bovine tissues were administered intravenously to rats after induction of disease. Fibronectin in glomeruli was quantitated to assess the ability of decorin to block matrix deposition in glomerulonephritis. Figure 2 shows that four or six daily injections of recombinant decorin or bovine decorin suppress fibronectin deposition to a level not significantly different from normal. Phosphate-buffered saline and three other solutions were used by themselves as controls. A cartilage proteoglycan, aggrecan, was used as a negatively charged molecule to control for the glycosaminoglycan chain of decorin. Aggrecan does not inhibit TGF- $\beta$  (ref. 6). Ovalbumin, which is approximately the same size as decorin, was used to control for the decorin core protein. Bovine serum albumin was used as an unrelated protein control. The proteoglycan and protein controls were indistinguishable from glomerulonephritic rats injected with buffer alone (Fig. 2). Two injections of decorin given early in the disease process also had no effect on fibronectin accumulation.

Next, glomeruli were stained with antibodies that detect fibronectin containing extra domain A (EDA<sup>+</sup> fibronectin) and tenascin, because these matrix components are specifically induced by TGF- $\beta$  (refs 7-9). EDA<sup>+</sup> fibronectin and tenascin, present only in trace amounts in glomeruli of normal rats, were

greatly increased in nephritic glomeruli from both untreated rats and rats treated with two injections of decorin (Fig. 3). But, four to six injections of decorin greatly reduced the deposition of these two markers of TGF- $\beta$  activity (Fig. 3). Histological analysis of periodic acid-Schiff-stained sections of the same kidneys showed a comparable effect of decorin in preventing glomerular extracellular matrix deposition and the associated glomerular enlargement that occurs in the disease (not shown).

Decorin given for 4 to 6 days also suppressed the development of proteinuria. The ratio of urinary protein concentration to urinary creatinine concentration is given because this ratio controls for differences in urinary volume<sup>10</sup>: mean values for normals, disease control, low-dose decorin (2 injections) and high-dose decorin-treated rats (4 to 6 injections) were 0.7, 15.2, 14.8 and 2.4, respectively ( $P < 0.01$  for high-dose decorin rats compared with disease controls).

Accumulation of extracellular matrix in glomerulonephritis is caused by overproduction of TGF- $\beta$  (refs 1, 2). Our results demonstrate a dramatic effect of decorin in preventing the deposition of extracellular matrix in injured glomeruli from nephritic rats, showing that decorin suppresses TGF- $\beta$  *in vivo*.

TGF- $\beta$ 1 is raised in glomerulonephritis and in many fibrotic diseases<sup>11</sup>; TGF- $\beta$ 2 is elevated in the eye of humans with proliferative vitreoretinopathy and in the skin in systemic sclerosis<sup>12,13</sup>. If different TGF- $\beta$  isoforms are involved in different diseases, then the ability of decorin to inhibit TGF- $\beta$ 1, 2 and 3 (Fig. 1) may be a therapeutic advantage.

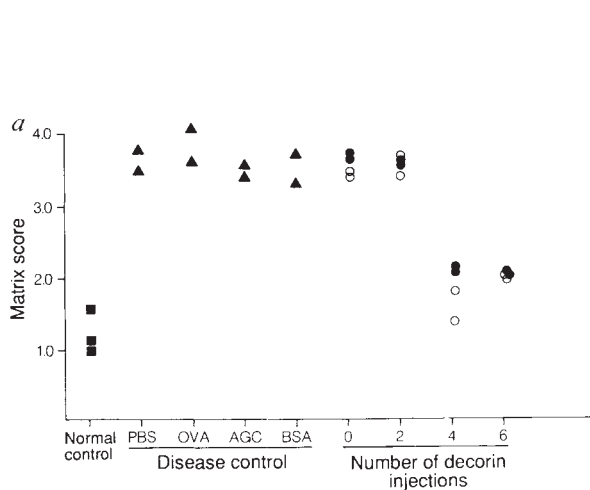
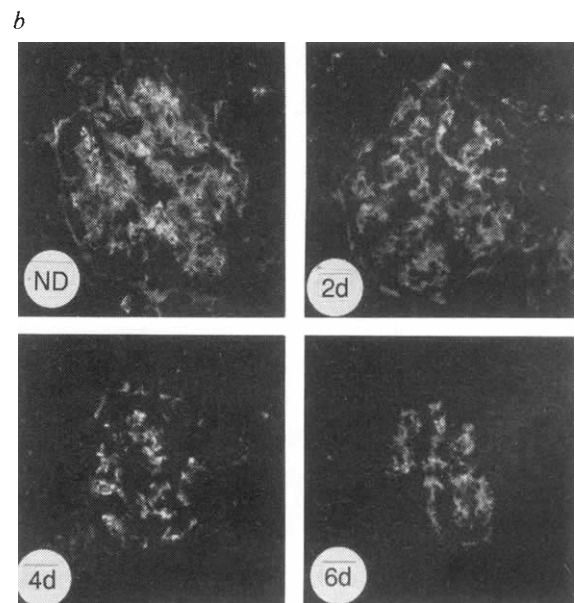


FIG. 2 Deposition of fibronectin in glomeruli of decorin-treated glomerulonephritic rats. *a*, Quantitation of fibronectin staining. Rats treated with four or six injections of decorin had significantly less fibronectin deposited in glomeruli than did the disease controls or rats that received no or two injections of decorin ( $P < 0.01$ ). ■, Normal control; ▲, disease control; ●, human decorin; ○, bovine decorin. *b*, Immunofluorescence micrographs of glomeruli from glomerulonephritic rats stained with anti-fibronectin antibody. Rats were treated with no decorin (ND) (that is, PBS alone) or decorin for two days (2 d), four days (4 d) or six days (6 d). The maximum standard error of the mean for glomerular scores in any animal was 0.16, indicating low interglomerular variability. Matrix scores of the illustrated glomeruli are: ND, 3.5; 2d, 3.5; 4d, 2.0; 6d, 2.0. Photographs were taken under identical conditions with exposure times of 40 s. Magnification, 340 $\times$ . METHODS. Glomerulonephritis was induced in Sprague-Dawley rats (4-6 weeks old) by intravenous injection of antithymocyte serum<sup>1</sup>. Recombinant human decorin and bovine decorin were tested in separate but identical experiments as follows. One hour after antithymocyte injection and then daily for 6 d, eight rats received a 0.5-ml intravenous injection of PBS or decorin (0.9 mg ml<sup>-1</sup> in PBS) as follows: group 1 (decorin, 2 d; PBS, 4 d); group 2 (decorin, 4 d; PBS, 2 d); group 3 (decorin, 6 d); group 4 (PBS, 6 d).



Each group contained 2 rats. In a third experiment, 8 rats (2 rats per group) received antithymocyte serum followed by either 0.5 ml PBS alone or PBS containing aggrecan (AGC), ovalbumin (OVA) or BSA. Four normal control rats received PBS instead of antithymocyte serum. Rats were restrained but not anaesthetized during injection. On day 7, a 24-h urine was collected from all rats. Urinary protein and creatinine measurement, sacrifice and tissue preparations were as described<sup>1</sup>. Fluorescein isothiocyanate-conjugated sheep anti-human fibronectin was from The Binding Site Inc. (San Diego). Immunofluorescence staining for fibronectin was scored blind using a scale of 0 (no staining) to 4 (strong staining) in 20 randomly selected glomeruli from each animal. Group differences were analysed by *t*-test. Recombinant human decorin was prepared from culture medium of a Chinese hamster ovary cell line, clone 62 (ref. 3). Bovine decorin was isolated from articular cartilage by extraction for 24 h with 4 M guanidine-HCl and protease inhibitors. Both decorins were purified by ion exchange on Q-Sepharose and octyl-Sepharose, as modified from ref. 17. After elution and dialysis against PBS, pH 7.4, decorin concentration was adjusted to 0.9 mg ml<sup>-1</sup>. The carbohydrate content of aggrecan was the same as that of decorin for injection solutions. Ovalbumin and BSA were treated with 4 M guanidine-HCl, dialysed, and adjusted to 0.9 mg protein per ml.

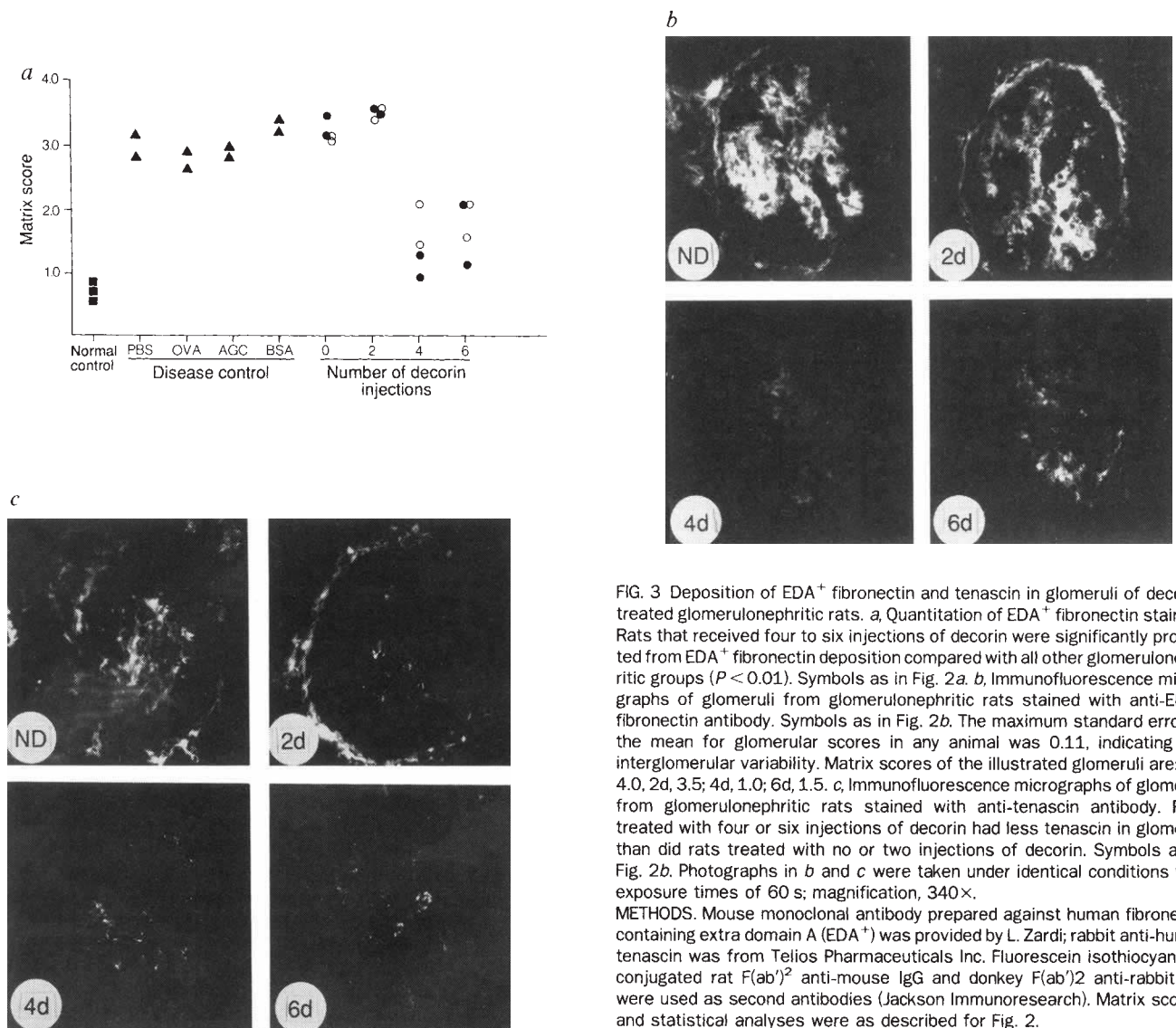


FIG. 3 Deposition of EDA<sup>+</sup> fibronectin and tenascin in glomeruli of decorin-treated glomerulonephritic rats. *a*, Quantitation of EDA<sup>+</sup> fibronectin staining. Rats that received four to six injections of decorin were significantly protected from EDA<sup>+</sup> fibronectin deposition compared with all other glomerulonephritic groups ( $P < 0.01$ ). Symbols as in Fig. 2*a*. *b*, Immunofluorescence micrographs of glomeruli from glomerulonephritic rats stained with anti-EDA<sup>+</sup> fibronectin antibody. Symbols as in Fig. 2*b*. The maximum standard error of the mean for glomerular scores in any animal was 0.11, indicating low interglomerular variability. Matrix scores of the illustrated glomeruli are: ND, 4.0; 2d, 3.5; 4d, 1.0; 6d, 1.5. *c*, Immunofluorescence micrographs of glomeruli from glomerulonephritic rats stained with anti-tenascin antibody. Rats treated with four or six injections of decorin had less tenascin in glomeruli than did rats treated with no or two injections of decorin. Symbols as in Fig. 2*b*. Photographs in *b* and *c* were taken under identical conditions with exposure times of 60 s; magnification, 340 $\times$ .

**METHODS.** Mouse monoclonal antibody prepared against human fibronectin containing extra domain A (EDA<sup>+</sup>) was provided by L. Zardi; rabbit anti-human tenascin was from Telios Pharmaceuticals Inc. Fluorescein isothiocyanate-conjugated rat F(ab')<sub>2</sub> anti-mouse IgG and donkey F(ab')<sub>2</sub> anti-rabbit IgG were used as second antibodies (Jackson ImmunoResearch). Matrix scoring and statistical analyses were as described for Fig. 2.

We used several markers to assess the effects of decorin on matrix accumulation and kidney function. Fibronectin is an abundant protein in the normal mesangial matrix of the rat and human and is greatly increased in humans with mesangial proliferative glomerulonephritis<sup>6,14-16</sup>. Using fibronectin staining as a marker of matrix, it was found that decorin significantly prevented the acute build-up of matrix that occurs by day 7 in the rat glomerulonephritis model. EDA<sup>+</sup> fibronectin and tenascin were more specific markers for TGF- $\beta$  activity and matrix accumulation, because these proteins are almost undetectable in normal adult rat kidney and because they are strongly induced by TGF- $\beta$ . Both markers were increased in the nephritic glomeruli and both were suppressed by decorin injections. A parallel reduction in proteinuria, a widely used indicator of glomerular damage in humans<sup>10</sup>, also showed that the effect on matrix build-up was therapeutic.

Suppressing TGF- $\beta$  activity had no detectable deleterious effects over the observation period; animals remained normal in appearance and behaviour. No toxic changes were seen in the kidneys, although radiolabelled decorin accumulates in the kidney, liver and lung (N.A.N. and W.A.B., unpublished results). Although they are probably critical during embryogenesis, TGF- $\beta$ s may be dispensable in the adult organism.

The diseases associated with overproduction of TGF- $\beta$  are chronic, and TGF- $\beta$  may perpetuate the disease process through

continued induction of its own production in cells at the site of injury<sup>11</sup>. In this regard it is interesting that administration of decorin during the first two days of disease has no effect. After tissue injury, TGF- $\beta$  is released from degranulating platelets and autoinduces TGF- $\beta$  production by local cells. In our model, increased expression of TGF- $\beta$  by glomerular cells is not detectable until 48 h after injury. Thus decorin may be effective only when TGF- $\beta$  is being produced by resident cells later in the disease process. Support for this idea comes from the observation that when decorin is given on days 3 and 4 or on days 4 and 5 only, excessive matrix deposition is suppressed (W.A.B. and N.A.N., unpublished results).

Decorin may be particularly suitable for the treatment of kidney, lung and liver disease because it has a propensity to accumulate in these organs after intravenous administration. Furthermore, decorin is a natural human compound which can be produced as a recombinant molecule and is therefore not likely to be immunogenic. Thus decorin offers a possible route to treatment for the many conditions associated with overproduction of TGF- $\beta$  for which there are currently few or no effective therapies. □

Received 16 June; accepted 9 October 1992.

- Okuda, S., Languino, L. R., Ruoslahti, E. & Border, W. A. *J. clin. Invest.* **86**, 453-462 (1990).
- Border, W. A., Okuda, S., Languino, L. R., Sporn, M. B. & Ruoslahti, E. *Nature* **346**, 371-374 (1990).



3. Yamaguchi, Y., Mann, D. M. & Ruoslahti, E. *Nature* **346**, 281–284 (1990).
4. Border, W. A., Okuda, S., Languino, L. R. & Ruoslahti, E. *Kidney Int.* **37**, 289–295 (1990).
5. Roberts, A. B. & Sporn, M. B. in *The Handbook of Experimental Pharmacology* (ed Sporn, M. B. & Roberts, A. B.) 419–472 (Springer, Heidelberg, 1989).
6. Oomura, A., Nakamura, T., Arakawa, M., Ooshima, A. & Isemura, M. *Virchows Arch. A. Pathol. Anat.* **415**, 151–159 (1989).
7. Balza, E., Borsi, L., Allemanni, G. & Zardi, L. *FEBS Lett.* **228**, 42–44 (1988).
8. Wang, G., Cohen, D. S., Palmer, E. & Sheppard, D. *J. Biol. Chem.* **266**, 15598–15601 (1991).
9. Pearson, C. A., Pearson, D., Shibahara, S., Hofsteenge, J. & Chiquet-Ehrismann, R. *EMBO J.* **7**, 2977–2981 (1988).
10. Ginsberg, J. M., Chang, B. S., Matarese, R. A. & Garella, S. *New Engl. J. Med.* **309**, 1543–1550 (1983).

11. Border, W. A. & Ruoslahti, E. *J. clin. Invest.* **90**, 1–7 (1992).
12. Connor, T. J. *et al. J. clin. Invest.* **83**, 1661–1666 (1989).
13. Kulozik, M., Hogg, A., Lankat-Büttgerit, B. & Krieg, T. *J. clin. Invest.* **86**, 917–922 (1990).
14. Courtoy, P. J., Kanwar, Y. S., Hynes, R. O. & Farquar, M. G. *J. Cell Biol.* **87**, 691–696 (1980).
15. Courtoy, P. J., Timpl, R. & Farquhar, M. G. *J. Histochem. Cytochem.* **30**, 874–886 (1982).
16. Border, W. A. *Kidney Int.* **34**, 419–434 (1988).
17. Choi, H. U., Johnson, T. L., Pal, S., Tang, L.-H. & Rosenberg, L. *J. Biol. Chem.* **264**, 2876–2884 (1989).

ACKNOWLEDGEMENTS. This work was supported by grants from the NIDDK (W.A.B.), the NCI and Cancer Center Support (E.R.) and from Telios Pharmaceuticals Inc. (W.A.B. and E.R.).

## Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle

Hwai-Chen Guo\*, Theodore S. Jardetzky\*,  
Thomas P. J. Garrett†, William S. Lane‡,  
Jack L. Strominger\* & Don C. Wiley†

\*Department of Biochemistry and Molecular Biology and

†Howard Hughes Medical Institute, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA

‡Harvard Microchemistry Facility, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA

WE report here the determination and refinement to 1.9 Å resolution by X-ray cryo-crystallography the structure of HLA-Aw68. The averaged image from the collection of bound, endogenous peptides clearly shows the atomic structure at the first three and last two amino acids in the peptides but no connected electron density in between. This suggests that bound peptides, held at both ends, take alternative pathways and could be of different lengths by bulging out in the middle. Peptides eluted from HLA-Aw68 include peptides of 9, 10 and 11 amino acids, a direct indication of the length heterogeneity of tightly bound peptides. Peptide sequencing shows relatively conserved 'anchor' residues at position 2 and the carboxy-terminal residue. Conserved binding sites for the peptide N and C termini at the ends of the class I major histocompatibility complex binding groove are apparently dominant in producing the long half-lives of peptide binding and the peptide-dependent stabilization of the class I molecule's structure.

Class I histocompatibility glycoproteins bind peptides from the cytoplasm of cells and 'present' them at the cell surface for surveillance by cytotoxic T lymphocytes (CTL)<sup>1</sup>. Foreign peptides, such as those from viral infections, complexed with major

histocompatibility complex (MHC) molecules present a novel antigenic surface to CTL, and are therefore recognized, stimulating CTL to kill the invaded cell. Peptide binding must be extremely 'tight', with a long half-life, in order to resist dissociation at the cell surface where the free peptide concentration is essentially near zero. Studies of peptides eluted from class I MHC molecules show that nonamers (that is, peptides that are nine amino acids long) are preferred<sup>2–7</sup> and that peptides from each MHC allele have a sequence 'motif' with certain positions ('anchors') conserved or nearly so<sup>5</sup> (reviewed in ref. 8).

The structure of HLA-Aw68, previously determined at 2.6 Å resolution<sup>9</sup>, has been determined and refined to 1.9 Å resolution from X-ray diffraction data collected from a single crystal frozen at –165 °C (see Fig. 1 legend). At 1.9 Å resolution, the electron density of the bound peptides is very readily interpreted at both ends of the HLA binding groove (Fig. 1a) as a atomic model of the N-terminal and C-terminal parts of bound peptides. Models of Glu-Val-Ala and Ala-Arg have been built into the first three (P1, P2, P3) and last two (PC-1, PC carboxy-terminal) positions of the peptide, respectively (based on eluted sequences). The real-space correlation coefficients for the main chain, a measure of how well the peptide atomic model fits the ( $F_o - F_c$ ) difference electron density<sup>10</sup>, are reasonably good for P1 (0.9), P2 (0.89) and PC (0.92), and comparable to that for the whole protein model (0.93) fit to the  $2F_o - F_c$  electron density map. By contrast, the middle part of the peptide density is sparse and unconnected (Fig. 1b, c), presumably representing an average of diverse conformations and/or lengths. The peptide model at the P3 and PC-1 positions, which might be positioned differently for various conformations or lengths in the peptide's centre, have correspondingly lower real-space correlation coefficients, 0.75 for P3 and 0.83 for PC-1.

The collection of peptides bound to HLA-Aw68 makes the same atomic contacts from the peptide's main chain and termini to a set of conserved residues at both ends of the binding cleft (Fig. 2), as observed from a single influenza virus peptide to HLA-Aw68<sup>11</sup> and from a collection of endogenous, nonameric

TABLE 1 Self-peptide sequences from HLA-Aw68

| Cycle | Peptide sequence |   |   |   |   |   |   |   |   |    |    | $M_{calc}$ | $M_{obs}$   | Homologous protein          |
|-------|------------------|---|---|---|---|---|---|---|---|----|----|------------|-------------|-----------------------------|
|       | 1                | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |            |             |                             |
|       | E                | V | A | P | P | E | Y | H | R |    |    | 1,097.2    | 1,097 ± 1   | —                           |
|       | A                | V | A | A | V | A | A | R | R |    |    | 884.0      | 884 ± 1     | —                           |
|       | D                | V | F | R | D | P | A | L | K |    |    | 1,060.2    | 1,060 ± 1   | Ribosomal 60S homologue     |
|       | E                | V | A | P | P | E | Y | H | R | K  |    | 1,225.4    | 1,225 ± 0.5 | —                           |
|       | E                | V | I | L | I | D | P | F | H | K  |    | 1,210.4    | 1,213 ± 4   | —                           |
|       | T                | V | F | D | A | K | R | L | I | G  | R  | 1,275.5    | ND          | Human Hsp70 protein B/Hsc70 |
|       | K                | T | G | G | P | I | Y | K | R |    |    |            |             | Influenza Np 91–99          |

Peptides were isolated, HPLC-purified and then microsequenced using an ABI model 477A protein sequencer as described in ref. 7.  $M_{calc}$ , calculated mass of peptides in daltons;  $M_{obs}$ , observed mass determined using either a Finnigan TSQ700 or LaserMat mass spectrometer; ND, not determined; Np peptide from refs 11 and 18. P2 and the C-terminal residues (boxed) are discussed in the text. The self peptides sequenced have a minimum occupancy relative to HLA-Aw68 from 0.12 to 0.4% based on initial amino-acid yields from sequencing and the starting amount of HLA-Aw68 (20 nanomol). (The residues at P1 and P3 are also consistent with the HLA-Aw68 structure as described here and in the accompanying paper<sup>11</sup>.) Asp and Glu at P1 could contact Arg 62 and non-polar residues at P3; Phe, Ile and Ala face a non-polar pocket similar to that described for HLA-B27 at P3 (ref. 12).

FIG. 1 Endogenous peptides bound to HLA-Aw68. *a*, Side view of the electron density within 1.8 Å of the first three and last two residues of the collection of endogenous peptides bound to HLA-Aw68. The 1.8 Å cutoff only truncated density at the peptide's centre, as shown below. *b*, Side view as in *a*, but including unconnected density near the peptide's middle. *c*, Top view, including unconnected density near the peptide's middle. (The  $\alpha_1$   $\alpha$ -helix runs from left to right. An atomic model of Glu-Val-Ala is included in the left side (N-terminal) density and Ala-Arg in the right side (C-terminal) density.) Figure generated with FRODO<sup>26</sup>. Maps are calculated with coefficients ( $F_o - F_c$ ) in the resolution range 12–1.9 Å and contoured at 2.3  $\sigma$ , where  $F_o$  is the observed structure factor amplitudes and  $F_c$  are amplitudes calculated from the atomic model (described below). To avoid bias, no atomic model for bound peptides was included in the refinement of the protein, or to generate the electron density maps shown in this report.

**METHODS.** 1.9 Å X-ray diffraction data were collected from one crystal frozen in a film of buffer<sup>27</sup> containing 20% glycerol at –165 °C on a Xentronics Detector<sup>28</sup> and processed with BUDDHA<sup>29</sup>,  $R_{\text{merge}} = 7.3\%$ . The structure was determined by molecular replacement using the 2.6 Å resolution room temperature coordinates<sup>9</sup> and refined with XPLOR<sup>30</sup> (details to be published elsewhere). The current atomic model of 2,998 non-hydrogen protein atoms and 186 water molecules (no peptide included) has an  $R$ -factor = 21.8% ( $F > 3\sigma$  6–1.9 Å),  $R_{\text{free}}$  = (ref. 31) 27.7% ( $F > 0, 6$ –1.9 Å), r.m.s.  $\Delta_{\text{bonds}} = 0.017$  Å; r.m.s.  $\Delta_{\text{angles}} = 3.2^\circ$ .

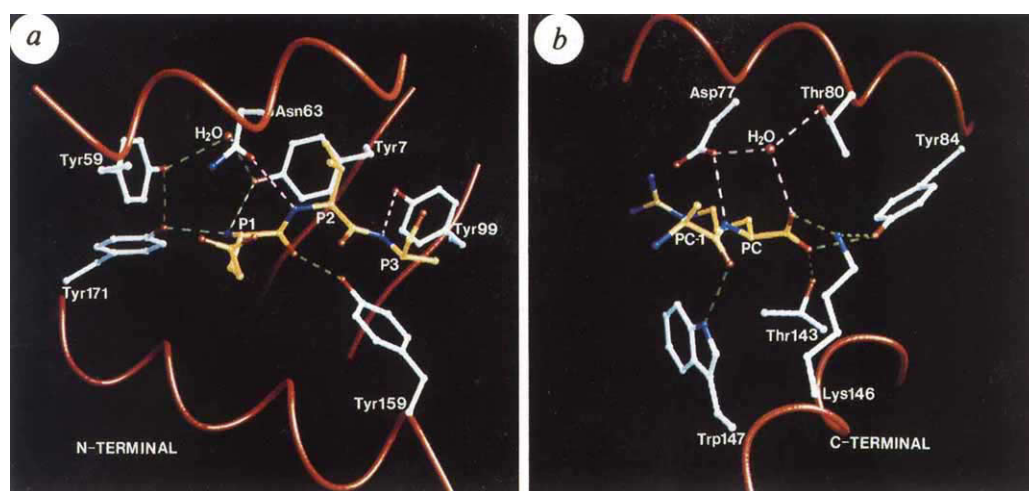
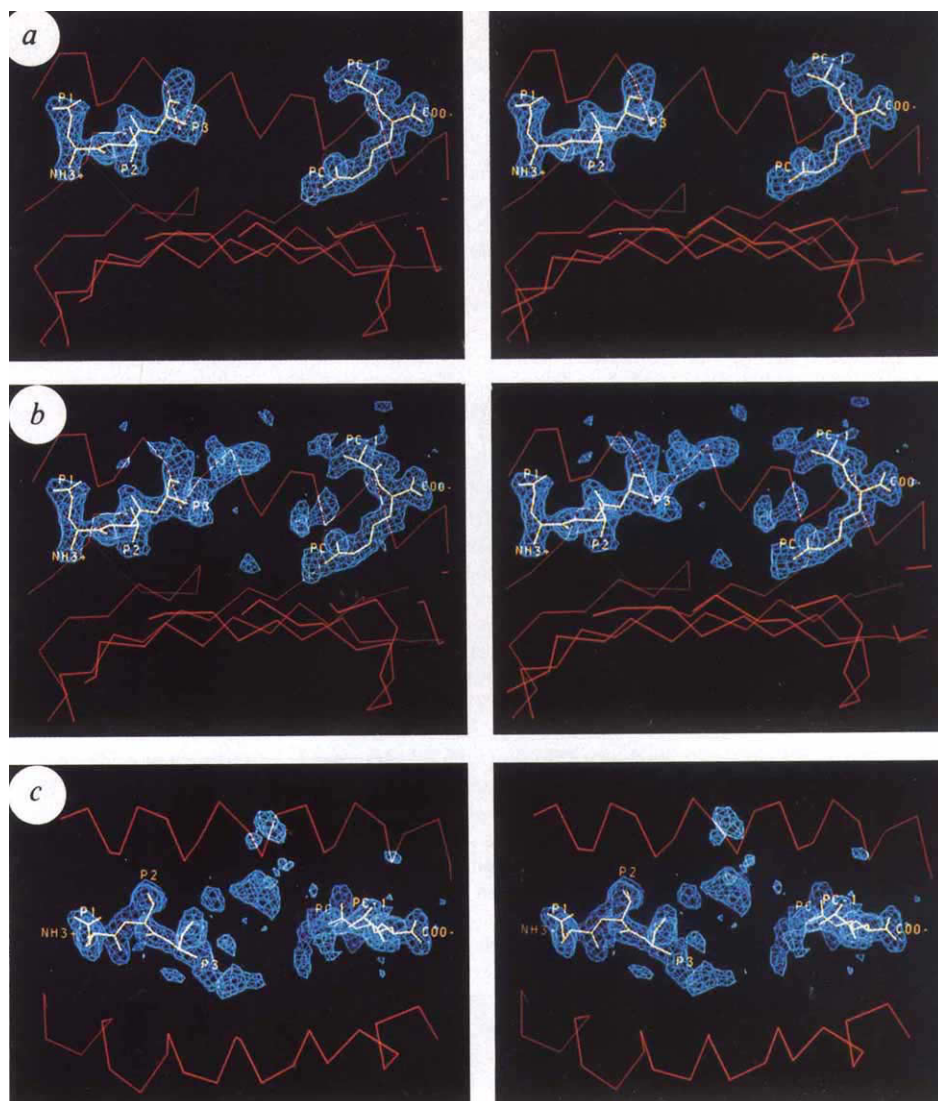


FIG. 2 A detailed view of the peptide model and hydrogen-bonding network at the N (*a*) and C (*b*) termini of the collection of endogenous peptides bound to HLA-Aw68. *a*, Potential hydrogen bonds (green dotted lines to conserved residues, purple to others) (distances less than 3.2 Å) between the peptide N terminus and HLA-Aw68. *b*, Potential hydrogen bonds between the peptide's C terminus and HLA-Aw68. HLA-Aw68  $\alpha$ -carbon trace in red; bonds coloured by atom type: yellow, peptide carbon; white, HLA-carbon; blue, nitrogen; red, oxygen. Figure generated with HYDRASTER, modified by S. Watowich and L. Gross from programs written by D. Bacon and W. Anderson (RASTER3D) and R. Hubbard (HYDRA). Figure view chosen for comparison with Fig. 5 of ref. 13.



peptides to HLA-B27<sup>7,12,13</sup>. In all three cases, the peptides' termini link together conserved residues on the sides (Tyr 59, Tyr 159, Tyr 171, Tyr 84, Thr 143, Lys 146, Trp 147) and bottom (Tyr 7, Tyr 123) of the peptide binding groove in a network of hydrogen-bonds (green in Fig. 2). This concentration of interactions at the peptides' ends not only holds the peptides in the groove, but seem to explain the peptide-dependent stabilization of the HLA structure<sup>14</sup>. Similar contacts have also been observed for two peptides bound to murine H-2K<sup>b</sup> (refs 15, 16).

The absence of connected density corresponding to the middle of the collection of peptides bound to HLA-Aw68 (P4 to PC-2) might be interpreted as an indication that short, nonameric peptides adopt different conformations as they extend from one end to the other. The suggestion that a single peptide might adopt different conformations when bound to MHC molecules<sup>17</sup> is not supported by the crystallographic structures of a single influenza virus peptide (Np 91-99) bound to HLA-Aw68 (ref. 11) and two other peptides bound to murine H-2K<sup>b</sup> (refs 15, 16), which show a unique conformation for those bound octamers and nonamers. In addition, the electron density corresponding to a collection of endogenous peptide nonamers<sup>7</sup> bound to crystalline HLA-B27 showed one pathway of clear density at high resolution (2.1 Å), readily interpretable as a common conformation for the collection of bound nonameric peptides<sup>13</sup>.

These observations suggest that the electron density at the centre of the collection of peptides bound to HLA-Aw68 might be missing because the bound peptides are of different lengths and consequently bulge out of the binding site to different degrees. We acid-eluted, fractionated and sequenced endogenous peptides<sup>4,7</sup> bound to HLA-Aw68. Six self peptides were sequenced with lengths ranging from 9 to 11 amino acids (Table 1). These peptides represent a random, although small sampling of the self peptides bound to HLA-Aw68, based on high-performance liquid chromatography peak height and separation, indicating that length heterogeneity is abundant in this pool. Furthermore, there is a small amino-acid 'anchor' at P2 (Val or Thr) and a positively charged amino acid 'anchor' (Arg or Lys) at the C-terminal position (boxed in Table 1) in six peptides sequenced so far, and in the influenza virus peptide, Np 91-99, that is restricted to HLA-Aw68 (refs 11, 18). This argues that

all of these peptides are bound at both ends by their charged termini and the P2 and PC anchors. Models of Val and Arg fit the electron density at P2 and PC with side-chain correlation coefficients of 0.81 and 0.84, comparable to the 0.89 for the HLA-Aw68 side chains. Peptides from 8 to 13 amino acids long have been reported in tight complexes with class I molecules and those sequences indicate that the extra length 'inserts' between anchor residues as found here (refs 2, 19-23, and A. Bertolotti *et al.*, personal communication).

It is not clear what limits the size of 'bulges' *in vivo* (for example, transporters, proteasomes, proteolysis) or *in vitro*, although length variability might be expected to be smaller for cases where the peptide residues contacting the MHC molecule are near the peptide centre (like H-2K<sup>b</sup>, P5, and HLA-B27, P7). A large enough insertion might even prevent a T-cell receptor from approaching the MHC molecule's surface. In HLA-peptide complexes with nonameric peptides, more than one half of the peptide side chains are exposed, primarily in the middle of the sequence, to recognition by T cells<sup>11-13</sup>. Where longer peptides can bulge out of the binding site, as in HLA-Aw68, each 'bulged' position can probably accommodate many different amino acids, substantially increasing the diversity of sequences that can be presented.

Crystallographic and sequence data argue that interactions at the ends of peptides dominate in binding to HLA-Aw68. The conservation of the HLA sites for binding peptide termini suggests that the same substantial number of interactions for binding a peptide and for stabilizing the MHC molecule's structure (Fig. 2) will be found in all tight peptide/class I molecule complexes. 'Anchor' peptide residues bind in polymorphic pockets that differ among alleles<sup>9,11-13,24,25</sup> and so might be expected to supply variable amounts of stabilization both to the peptide-HLA interaction and to the HLA molecule's stability. Polymorphic pockets therefore presumably determine which peptide sequences can bind<sup>9,24</sup> in such a way that the peptides' ends will fit into both of the conserved peptide terminal binding sites. This suggests that short peptide termini joined by a long non-peptide linker could be designed to bind to HLA, achieving both long half-life, 'tight' binding and HLA structure stabilization. □

Received 4 August; accepted 2 October 1992.

1. Brodsky, F. M. & Guagliardi, L. E. *Rev. Immun.* **9**, 707-744 (1991).
2. Rötzschke, O. *et al. Nature* **348**, 252-254 (1990).
3. van Bleek, G. M. & Nathenson, S. G. *Nature* **348**, 213-216 (1990).
4. van Bleek, G. M. & Nathenson, S. G. *Proc. natn. Acad. Sci. U.S.A.* **88**, 11032-11036 (1991).
5. Falk, K., Rötzschke, O., Stevanović, S., Jung, G. & Rammensee, H.-G. *Nature* **351**, 290-296 (1991).
6. Hunt, D. F. *et al. Science* **255**, 1261-1263 (1992).
7. Jardetzky, T. S., Lane, W. S., Robinson, R. A., Madden, D. R. & Wiley, D. C. *Nature* **353**, 326-329 (1991).
8. Rötzschke, O. & Falk, K. *Immun. Today* **12**, 447-455 (1991).
9. Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L. & Wiley, D. C. *Nature* **342**, 692-696 (1989).
10. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. *Acta crystallogr.* **A47**, 110-119 (1991).
11. Silver, M. L., Guo, H.-C., Strominger, J. L. & Wiley, D. C. *Nature* **360**, 367-369 (1992).
12. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Nature* **353**, 321-325 (1991).
13. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Cell* **70**, 1035-1048 (1992).
14. Townsend, A. *et al. Nature* **340**, 443-448 (1989).
15. Zhang, W., Young, A. C. M., Imarai, M., Nathenson, S. G. & Sacchettini, J. C. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8403-8407 (1992).
16. Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. *Science* **257**, 919-927 (1992).
17. Peccoud, J., Dellabona, P., Allen, P., Benoist, C. & Mathis, D. *EMBO J.* **9**, 4215-4223 (1990).

18. Cerundolo, V., Tse, A. G. D., Salter, R. D., Parham, P. & Townsend, A. *Proc. R. Soc.* **244**, 169-177 (1991).
19. Robbins, P. A. *et al. J. Immun.* **143**, 4098-4103 (1989).
20. Silver, M. L., Parker, K. C. & Wiley, D. C. *Nature* **350**, 619-622 (1991).
21. Wei, M. L. & Cresswell, P. *Nature* **356**, 443-446 (1992).
22. Henderson, R. A. *et al. Science* **255**, 1264-1266 (1992).
23. Deres, K. *et al. Eur. J. Immun.* **22**, 1603-1608 (1992).
24. Saper, M. A., Bjorkman, P. J. & Wiley, D. C. *J. molec. Biol.* **219**, 277-319 (1991).
25. Bjorkman, P. J. *et al. Nature* **329**, 512-518 (1987).
26. Jones, T. A. *J. appl. Crystallogr.* **11**, 268-272 (1978).
27. Teng, T.-Y. *J. appl. Crystallogr.* **23**, 387-391 (1990).
28. Durbin, R. M. *et al. Science* **232**, 1127-1132 (1986).
29. Blum, M., Metcalf, P., Harrison, S. C. & Wiley, D. C. *J. appl. Crystallogr.* **20**, 235-242 (1987).
30. Brünger, A. T. *X-PLOR Manual* (Version 2.1) (Yale University, New Haven, 1990).
31. Brünger, A. T. *Nature* **355**, 472-475 (1992).

ACKNOWLEDGEMENTS. We acknowledge the earlier contributions made by M. Saper and P. Bjorkman to the determination of the structure of HLA-Aw68 to 2.6 Å resolution<sup>9</sup>. We thank J. Gorga for supervising the protein purifications, A. Haykov and K. Svenson for excellent technical assistance and R. Robinson for peptide separation H.-C.G. and T.S.J. are Cancer Research Institute fellows. J.L.S. acknowledges support from the NIH. D.C.W. is supported and T.P.J.G. was supported by the Howard Hughes Medical Institute.

# Atomic structure of a human MHC molecule presenting an influenza virus peptide

Michael L. Silver\*, Hwai-Chen Guo\*,  
Jack L. Strominger\* & Don C. Wiley†

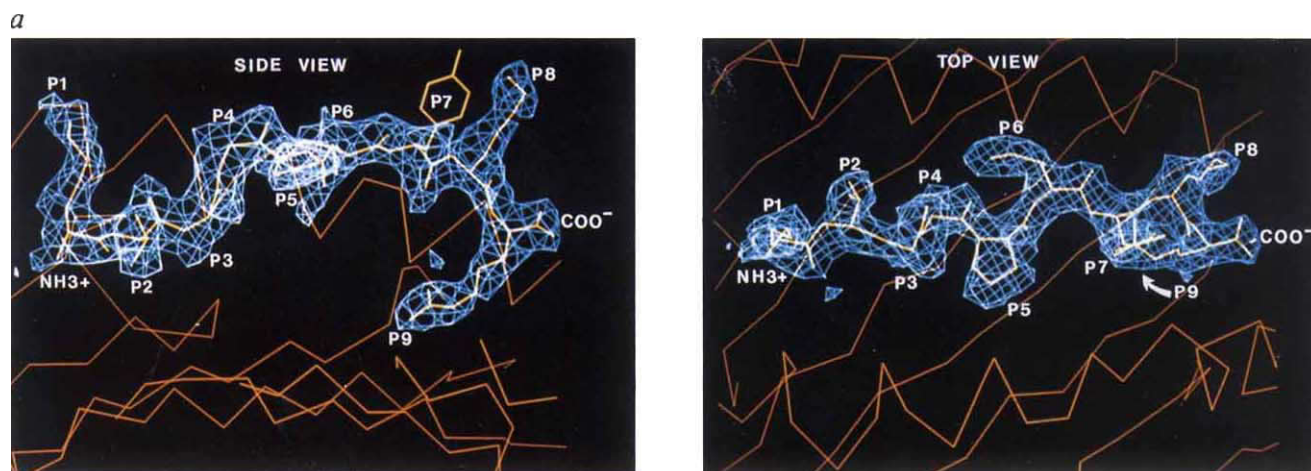
\* Department of Biochemistry and Molecular Biology and  
† Howard Hughes Medical Institute, Harvard University, 7 Divinity Avenue,  
Cambridge, Massachusetts 02138, USA

**INFECTION** by influenza virus results in the stimulation of cytotoxic T lymphocytes specific for killing virally infected cells<sup>1</sup>. Specificity is provided by clonally distributed, hypervariable T-cell receptors on cytotoxic T lymphocytes which react with peptide fragments that are derived from viral proteins expressed in the cytoplasm and 'presented' on the surface of infected cells, bound to class I histocompatibility glycoproteins<sup>2</sup>. Here we describe the structure of the complex between the human class I histocompatibility glycoprotein HLA-Aw68 and the influenza virus nucleoprotein peptide Np 91-99 as determined by X-ray cryocrystallography. Residues at both ends of the peptide are substantially buried in the peptide binding-site, whereas those in the middle of the peptide, P4 to P8, are predominantly exposed and could be recognized directly by T-cell receptors. The extended conformation of the bound viral peptide is remarkably similar to that of a

collection of endogenous peptides with a different sequence motif bound to another human allele, HLA-B27<sup>3-5</sup>. The structure defines in atomic detail the antigenic surface constructed of major histocompatibility complex and viral peptide atoms that is recognized by T-cell receptors.

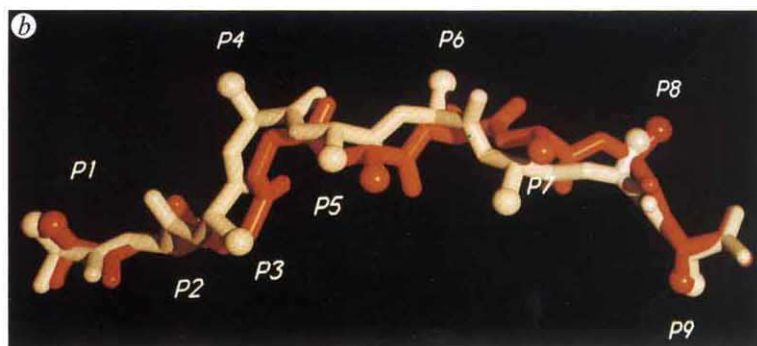
Crystals of HLA-peptide complex were grown from HLA-Aw68, isolated from human lymphoblastoid cells<sup>6,7</sup> and reconstituted *in vitro*<sup>8</sup> with the synthetic peptide Np 91-99 (Lys-Thr-Gly-Gly-Pro-Ile-Tyr-Lys-Arg)<sup>9</sup>. The structure was determined to 2.8 Å resolution from a single crystal frozen at -160 °C and refined to an *R*-factor of 18.5% with good geometry (see Fig. 1 legend). Electron density corresponding to Np 91-99 (Fig. 1a) was clearly interpretable with the exception of the side chain of tyrosine (P7) which appears to extend into solution and to be disordered.

The peptide conformation and mode of binding to the HLA groove are essentially the same for Np 91-99 binding to HLA-Aw68 as for multiple endogenous self-peptides bound to HLA-B27<sup>3-5</sup>. When the HLA-Aw68 and HLA-B27 molecules are superimposed, the atomic models of their bound peptides are nearly coincident (Fig. 1b). In both cases the peptide is extended with a kink at P3 and P4. Peptide side chains P2, P3 and P9 face into the major histocompatibility complex (MHC) groove, P1, P4, P8 face away towards the T-cell receptor (TCR), and P5 and P6, although accessible to a TCR, point back and forth across the top of the groove (Fig. 1a). The side chain at P7 points toward the TCR in the Aw68-Np 91-99 complex, but toward the groove in HLA-B27. Thus, the overall mode of



**FIG. 1** The structure of influenza virus peptide Np 91-99 as bound to HLA-Aw68. *a*, Top and side views of the 2.8 Å difference electron density with peptide model. The  $\alpha_1$  domain  $\alpha$ -helix runs from left to right. *b*, Comparison of the peptide conformations of Np 91-99 (red) bound to HLA-Aw68 and the model from a collection of endogenous peptides bound to HLA-B27 (white)<sup>3-5</sup>. The main-chain coordinates of HLA-Aw68 and HLA-B27 both without peptide coordinates were superimposed by FITATOM, a locally modified version of the program EXIFIT written by A. D. MacLachlan. The transformation matrix obtained was then used to superimpose the peptide models.

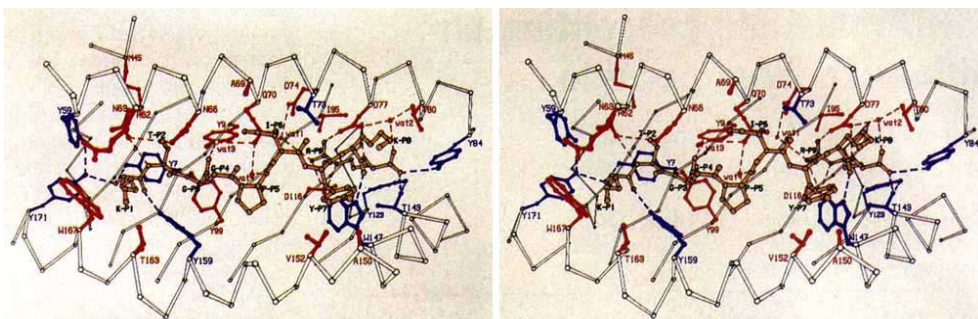
**METHODS.** The crystal structure of HLA-Aw68 complexed with Np 91-99 was determined by molecular replacement using 2.1 Å refined coordinates from a crystal of HLA-Aw68 complexed with a collection of endogenous peptides<sup>10</sup>. X-ray data to 2.8 Å were collected from a single crystal, frozen in a film of buffer in a wire loop<sup>14</sup> at -160 °C, using a Xentronics detector<sup>15</sup> and processed with BUDDHA software<sup>16</sup>. Subsequent calculations used AGROVATA and ROTAVATA<sup>17</sup> ( $R_{\text{merge}}$  = 10% to 2.8 Å) and XPLOR<sup>18</sup>. Initial molecular replacement solution plus positional minimization of HLA atoms revealed clear density for peptide main chain and four of the seven side chains (Thr P2, Pro P5, Ile P6, Arg P9) in an  $F_o - F_c$  map. A peptide model was built into the clear density. Subsequent cycles of refinement brought up electron density



interpretable for two of the remaining three side chains (Lys P1, Lys P8) (r.m.s. deviation in bonds = 0.016 Å, r.m.s. deviation in angles = 3.56°). Np 89-101 had been identified as an influenza epitope restricted to HLA-Aw68<sup>9</sup>. Np 91-99 was chosen from a series of truncated analogues because it produced the maximum in yield in *in vitro* complex formation<sup>19</sup>. *a*, Generated with FRODO<sup>20</sup> and *b*, with RASTER3D, a program written by D. Bacon and W. F. Anderson.



FIG. 2 Peptide binding interactions. Peptide in orange. HLA residue side chains containing contacts within 4 Å of peptide are shown. Conserved human class I MHC residues in blue (143 is Ser and 147 is Leu in one sequence). Variable MHC residues and water molecules in red. Dashed lines show potential hydrogen bonds from Np 91–99 to HLA-Aw68, blue dashes to conserved class I MHC residues, red dashes to variable class I MHC residues either directly, or bridged by water. Selection of conserved and polymorphic residues based on ref. 21. MHC residues of low variability are 63, 99, 150, 152 and 167. Stereo figure generated with ORTEP<sup>22</sup>.



peptide binding is extremely similar, although HLA-Aw68 and HLA-B27 differ by 12 amino acids in the peptide-binding site and the sequence motif of peptides that bind to the two alleles also differs<sup>4,10</sup>.

Atomic contacts at the two ends of the peptide predominate in binding Np 91–99 to HLA-Aw68. Residues P1, P2, P3 and P9 are substantially buried in the antigen-binding cleft. In contrast, the central peptide residues P4 to P7 make no hydrogen bonds to the HLA molecule and P4 and P5 do not even make van der Waal contacts (Fig. 2). Eleven MHC residues contact the first two peptide amino acids 'P1 and P2 and eight other MHC residues contact the last peptide amino acid P9, whereas all the rest of the peptide, P3 to P8, is in contact with only 10 MHC residues. In the centre of the peptide where direct contact to HLA is sparse, water molecules provide a hydrogen-bonding bridge to HLA (Fig. 2).

Conserved side chains of the HLA molecule hydrogen bond to main-chain polar atoms at both ends of the peptide: HLA residues Tyr 7 and Tyr 171 to the peptide N terminus, Tyr 159 to the P1 carbonyl oxygen, Tyr 84 and Thr 143 to the C-terminal carboxylate, and Trp 147 to the P8 carbonyl oxygen (Fig. 2). Similar hydrogen-bond networks, constituting conserved peptide terminal binding sites, have been visualized in the complexes of multiple endogenous peptides bound to HLA-B27<sup>3–5</sup> and to HLA-Aw68 (Fig. 2 in ref. 10).

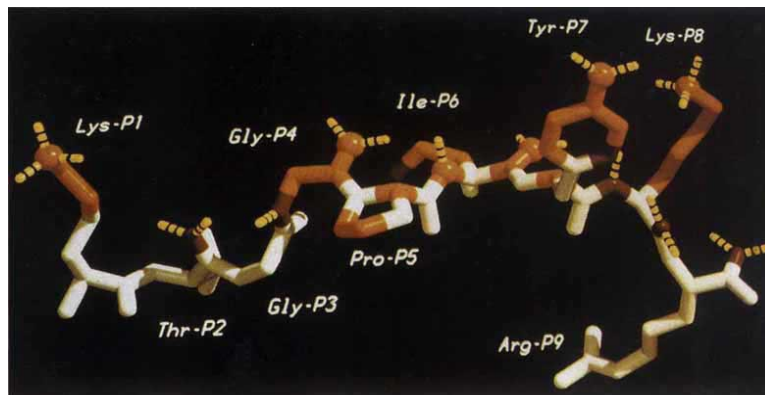
Peptide side chains Thr P2, and Arg P9 are entirely buried in binding cleft pockets composed of polymorphic residues. The side chain of threonine P2 is bound by atoms from five MHC residues (Tyr 7, Met 45, Asn 63, Asn 66, Tyr 99) in the shallow '45' pocket of HLA-Aw68. The side chain of arginine P9 reaches deep into the site, forming salt-bridged hydrogen bonds with the polymorphic residues Asp 77, and Asp 116 (Fig. 2) and van der Waal contacts with a total of six MHC residues (Asp 74, Asp 77, Met 95, Asp 116, Thr 143, Trp 147), four of which (74, 77, 95, 116) are polymorphic in the population. The contacts to Arg P9 are similar to those that Arg P9 would make in HLA-

B27 which shares many of the same polymorphic residues as Aw68 in the 'P9 end' of the binding site<sup>3</sup>. The binding of arginine P9 to the negatively charged pocket near Asp 74 was predicted from considering the HLA-Aw68 structure and making substituted peptides, as was the accessibility of Lys P8 to cytotoxic T lymphocyte (CTL) recognition<sup>9,11</sup>. If P3 (Gly) had a side chain it would point into a shallow non-polar pocket. In the accompanying paper<sup>10</sup> HLA-Aw68 is shown to prefer peptides with Val or Thr at P2 and Lys or Arg at P9, as well as a non-polar side chain at P3. That is consistent with the observations here where the two 'anchors' P2 and P9 of Np 91–99 are completely sequestered by MHC atoms.

Six of the nine peptide residues, P1, P4, P5, P6, P7 and P8 remain solvent accessible in the complex with HLA-Aw68 and could be recognized directly by the T cell (Fig. 3). Most of the accessible atoms are from the central residues of the peptide (red in Fig. 3). Six of the peptide's polar groups are accessible with the potential to form hydrogen bonds directly to a TCR, three from side chains Lys P1, Tyr P7 and Lys P8 and three from the main chain (yellow bonds in Fig. 3).

The structure of the complex between the influenza virus peptide Np 91–99 and the human class I histocompatibility glycoprotein HLA-Aw68 shows a class I molecule holding a peptide predominantly by its ends, stretched out so that most of its sequence can be 'read' by TCR (Fig. 4). This makes possible the detection of foreign gene products within a host cell by T-cell surveillance of class I MHC/peptide complexes. Six of the nine peptide residues could be contacted directly by TCR, a number sufficient to discriminate between a very large number of sequences. (A mechanism for presenting an even larger number of sequences, by insertion of additional peptide residues, is reported in the accompanying paper<sup>10</sup>.) No large conformational changes were found between HLA-Aw68 complexed with a single viral peptide, Np 91–99, and HLA-Aw68 complexed with multiple endogenous peptides<sup>10</sup>, suggesting in this case that peptide induced conformational changes of the HLA molecule

FIG. 3 Influenza virus nucleoprotein peptide Np 91–99 showing atoms accessible to solvent only (dark red) accessible to contact by TCR (orange), or buried by contact with HLA-Aw68 (white). Parts of residue P1 and P4–P8 are accessible to TCR. The solvent accessibility was determined using the program ACCESS written by M. D. Handschumacher and F. M. Richards with a 1.4 Å probe. TCR accessibility was approximated with a 1.9 Å probe representing an average van der Waals radius of TCR atoms. Yellow striped bands represented potential hydrogen bonds directly to TCR (on orange enlarged atoms) or to TCR through water molecules only (on red atoms). Figure generated with RASTER3D. (A total of 1,098 Å<sup>2</sup> of the solvent-accessible surface of the free Np 91–99 peptide is buried in the complex with HLA-Aw68 and 711 Å<sup>2</sup> of the HLA-Aw68 solvent-accessible surface is buried.)



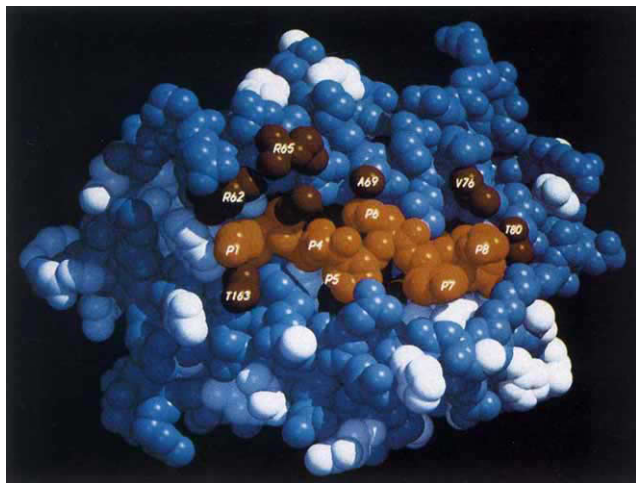


FIG. 4 The 'antigenic' surface composed of Np 91-99 peptide (orange) and MHC atoms (blue, conserved; red, polymorphic; light blue, not conserved or polymorphic). N-terminal of peptide is to the left,  $\alpha_1$  domain  $\alpha$ -helix top,  $\alpha_2$  domain  $\alpha$ -helix bottom. Figure generated with RASTER3D. (Of the 12 polymorphic residues facing into the binding site, 8 contact the peptide directly (9, 45, 66, 70, 74, 77, 95, 116) and four do not (67, 97, 114, 156), but 11 of the 12 (66 excluded) are nevertheless completely buried by the bound peptide. These polymorphic positions must therefore, as anticipated<sup>23</sup>, have their primary effect on T-cell recognition of HLA-Aw68 through the choice of peptides that can bind. Of the six polymorphic residues that face more directly toward solvent (62, 65, 69, 76, 80, 163), four also contact the peptide (62, 69, 80, 163) but all have atoms accessible to direct recognition by the TCR and therefore represent polymorphism recognizable by TCRs in the presence or absence of peptide.)

may not be a major factor in the creation of novel antigenic surfaces recognized by T cells. On the basis of the number of atomic contacts, Np 91-99 appears to be bound to HLA-Aw68 predominantly by two main features of the MHC molecule: (1) conserved MHC residues hydrogen bond to the peptide termini; (2) polymorphic MHC residues bury the two 'anchor' peptide side chains. Although both of these sets of interactions would also provide for the peptide-dependent stabilization of the MHC molecule, only the peptide termini binding sites are conserved in class I histocompatibility antigen sequences. The overall mode of peptide binding observed here seems to be a general mechanism for class I MHC presentation now visualized in three human alleles and one murine allele: HLA-B27<sup>3-5</sup>, HLA-Aw68<sup>10</sup>, HLA-A2<sup>12</sup> and H-2K<sup>b</sup> (refs 13, 24). □

22. Johnson, C. K. ORTEP Manual (Oak Ridge National Laboratory, Tennessee, 1965).

23. Bjorkman, P. J. *et al. Nature* **329**, 512-518 (1987).

24. Zhang, W., Young, A. C. M., Imai, M., Nathenson, S. G. & Sacchettini, J. C. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8403-8407 (1992).

ACKNOWLEDGEMENTS. We acknowledge the contributions made by T. P. J. Garrett, M. A. Saper, and P. J. Bjorkman to the determination of the structure of HLA-Aw68 to 2.6 Å resolution<sup>11</sup>. We thank J. Gorga for supervising the protein purifications and A. Haykov, K. Svenson and R. Crouse for technical assistance. H.-C.G. is a Cancer Research Institute fellow. J.L.S. and D.C.W. acknowledge support from the NIH. D.C.W. is supported by the Howard Hughes Medical Institute.

## The three-dimensional structure of an intact monoclonal antibody for canine lymphoma

Lisa J. Harris, Steven B. Larson, Karl W. Hasel\*, John Day, Aaron Greenwood & Alexander McPherson

Department of Biochemistry, University of California Riverside, Riverside, California 92521, USA

\* Immunopharmaceutics Inc., 11011 Via Frontera, San Diego, California 92127, USA

CRYSTAL structures of Fab antibody fragments determined by X-ray diffraction characteristically feature four-domain,  $\beta$ -barrel arrangements<sup>1-3</sup>. A human antibody Fc fragment has also been found to have four  $\beta$ -barrel domains<sup>4</sup>. The structures of a few intact antibodies have been solved<sup>5-8</sup>: in two myeloma proteins, the flexible hinge regions that connect the Fc to the Fab segments were deleted<sup>5,6</sup> so the molecules were non-functional, structurally restrained, T-shaped antibodies; a third antibody, Kol, had no hinge residues missing but the Fc region was sufficiently disordered that it was not possible to relate its disposition accurately with respect to the Fab components<sup>7,8</sup>. Here we report the structure at 3.5 Å resolution of an IgG2a antitumour monoclonal antibody which contains an intact hinge region and was solved in a triclinic crystal by molecular replacement using known Fc and Fab fragments. The antibody is asymmetric, reflecting its dynamic character. There are two local, apparently independent, dyads in the molecule. One relates the heavy chains in the Fc, the other relates the constant domains of the Fabs. The variable domains are not related by this 2-fold axis because of the different Fab elbow angles of 159° and 143°. The Fc has assumed an asymmetric, oblique orientation with respect to loosely tethered yet almost collinear Fabs. Our study enables the two antigen-binding segments as well as the Fc portion of a functional molecule to be visualized and illustrates the flexibility of these immune response proteins.

The specific murine antibody described here reacts with cells of canine lymphoma<sup>9</sup>, the most common haemopoietic tumour in the dog, which resembles human non-Hodgkin's lymphoma. This antibody can participate in antibody-dependent cellular cytotoxicity as well as complement-dependent cytotoxicity<sup>10</sup> and is used as an anticancer therapeutic<sup>11</sup> by veterinarians. The immunoglobulin crystallizes from a low concentration of polyethylene glycol at slightly alkaline pH (ref. 12).

The structure of the triclinic crystals, having one entire antibody as the asymmetric unit, was solved using the method of molecular replacement as implemented in the programs MERLOT<sup>13</sup> for rotation functions and XPLOR<sup>14,15</sup> for translation functions. By virtue of the triclinic cell there were no symmetry constraints on the molecule, immediately suggesting that the antibody has an asymmetric conformation. Molecular probe coordinates for Fab fragments and the Fc fragment were obtained from the Brookhaven Data Bank<sup>16</sup>. Sometimes cross-rotation<sup>17</sup> searches were done with probes representing one third of the asymmetric unit, and in other cases the search probes comprised only one sixth of the asymmetric unit. Convincing

Received 4 August; accepted 2 October 1992.

1. Townsend, A. & Bodmer, H. A. *Rev. Immun.* **7**, 601-624 (1989).
2. Brodsky, F. M. & Guagliardi, L. E. A. *Rev. Immun.* **9**, 707-744 (1991).
3. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Nature* **253**, 321-325 (1991).
4. Jardetzky, T. S., Lane, W. S., Robinson, R. A., Madden, D. R. & Wiley, D. C. *Nature* **253**, 326-329 (1991).
5. Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. *Cell* **70**, 1035-1048 (1992).
6. Turner, M. J. *et al. J. biol. Chem.* **250**, 4512-4519 (1975).
7. Bjorkman, P. J., Strominger, J. L. & Wiley, D. C. *J. molec. Biol.* **186**, 205-210 (1986).
8. Silver, M. L., Parker, K. C. & Wiley, D. C. *Nature* **350**, 619-622 (1991).
9. Cerundolo, V., Tse, A. G. D., Salter, R. D., Parham, P. & Townsend, A. *Proc. R. Soc.* **244**, 169-177 (1991).
10. Guo, H.-C., Jardetzky, T. S., Lane, W. S., Strominger, J. L. & Wiley, D. C. *Nature* **360**, 364-366 (1992).
11. Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L. & Wiley, D. C. *Nature* **342**, 692-696 (1989).
12. Saper, M. A., Bjorkman, P. J. & Wiley, D. C. *J. molec. Biol.* **219**, 277-319 (1991).
13. Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. *Science* **257**, 919-927 (1992).
14. Teng, T.-Y. *J. appl. Crystallogr.* **23**, 387-391 (1990).
15. Durbin, R. M. *et al. Science* **232**, 1127-1132 (1986).
16. Blum, M., Metcalf, P., Harrison, S. C. & Wiley, D. C. *J. appl. Crystallogr.* **20**, 235-242 (1987).
17. Fox, G. C. & Holmes, K. C. *Acta Crystallogr. A* **34**, 517 (1966).
18. Brünger, A. T. XPLOR (version 2.1) Yale University, New Haven, (1990).
19. Silver, M. L. thesis, Harvard University (1992).
20. Jones, T. A. *J. appl. Crystallogr.* **11**, 268-272 (1978).
21. Parham, P. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 4005-4009 (1988).



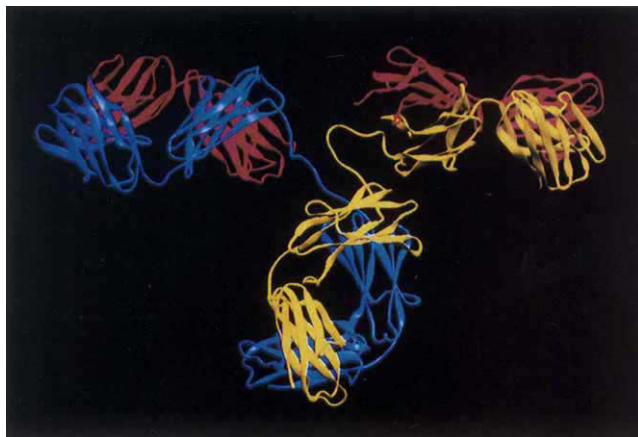


FIG. 1 Ribbon representation of the structure of the murine antibody against canine lymphoma determined by X-ray analysis of the triclinic crystals. The heavy chains are shown in yellow and blue, the light chains in red. The Fc stem of the molecule projects towards the viewer and assumes an asymmetric, oblique orientation with respect to the Fabs. This orientation illustrates the large difference in hinge angles of about 65° and 115°. The local dyad relating the heavy chains of the Fc is that dyad indicated by the primary solution of the self-rotation function. Fab2 is viewed along the axis through the switch peptides. Fab2 has an elbow angle of 143°, in contrast to Fab1, which has an elbow angle of 159°. Twenty-three residues in each heavy chain, comprising the hinge regions seen here, were built into the model with idealized geometry using both FRODO<sup>29</sup> and XPLOR<sup>14,15</sup>. These residues were missing from the fragment models taken from the Brookhaven Data Bank.

FIG. 2 *a*, Stereo diagram of the monoclonal antibody viewed perpendicular to the approximate 2-fold axis relating the constant domains of the Fabs. This dyad was that indicated by the secondary solution of the self-rotation function. Apparent here is the difference in the two elbow angles and the consequent failure of the variable domains to maintain this relationship. Also apparent in this view is the failure of the Fc dyad to intersect the 2-fold axis relating the constant domains of the Fabs. Both symmetry axes are apparently independent local dyads. *b*, Stereo diagram of the IgG2a antibody showing the region between the CH2 domains. In the human IgG1 Fc fragment<sup>4</sup>, carbohydrate was located in this area between the two CH2 domains and is probably in a similar location in this antibody. No attempt has yet been made to include the carbohydrate component in the model. It can also be seen here that the dyad axis of the Fc does not intersect the approximate long axis of the Fabs. Colour coding is the same as in Fig. 1.

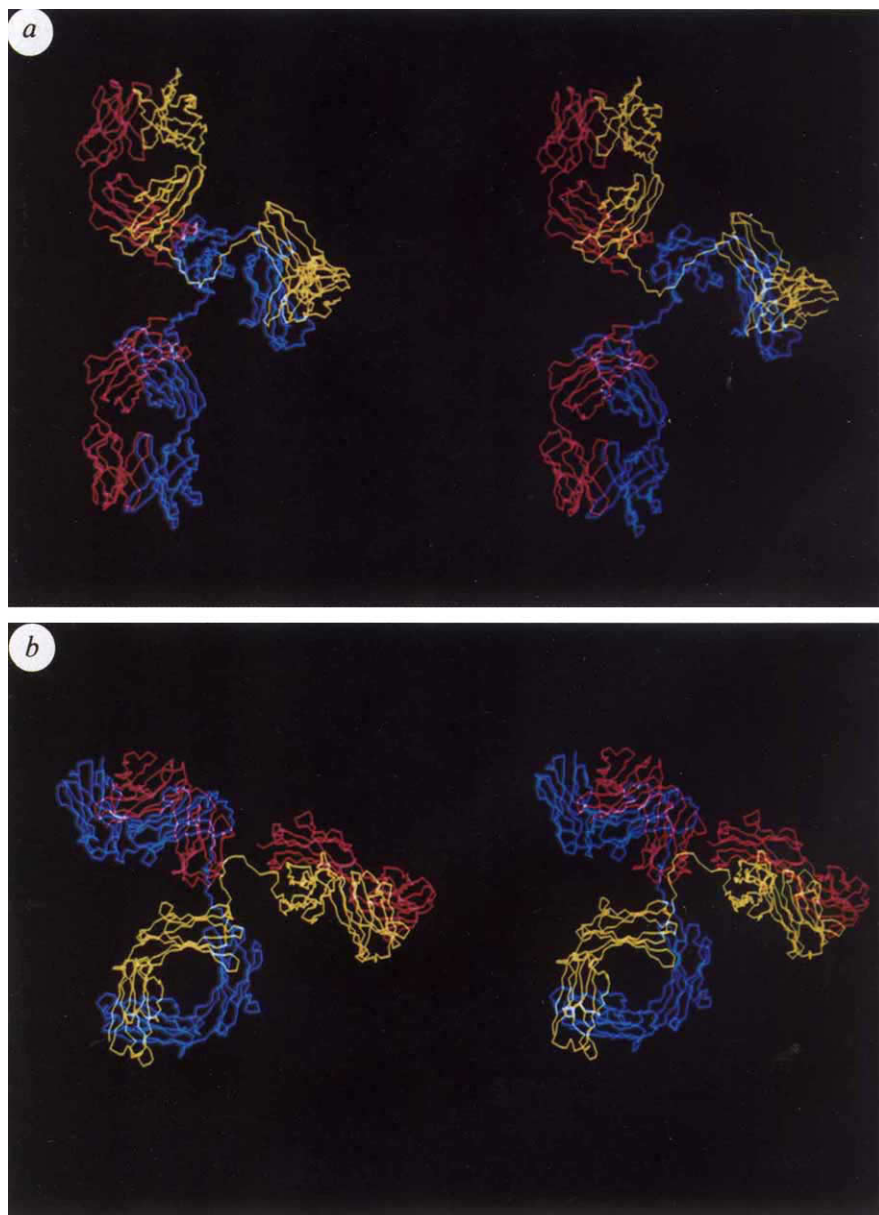


TABLE 1 Crystallographic data, structure solution and refinement

Crystal data: Space group  $P1$ ;  $a=66.39$  Å,  $b=77.34$  Å,  $c=101.42$  Å,  $\alpha=87.6^\circ$ ,  $\beta=92.6^\circ$ ,  $\gamma=97.5^\circ$ ,  $Z=1$ ; resolution  $\sim 3.0$  Å. Data collection: SDMS (Xuong-Hamlin) detectors, Rigaku Ru-200 source, frame size  $0.12^\circ$ , counting time 60–120 s. Total observations, 114,867; at 3.5 Å unique reflections, 24,808; 98.7% complete;  $R_{\text{sym}}=0.098$ . After  $F/\sigma=4.0$  cutoff, unique reflections at 3.5 Å = 20,964

| Operation                                                                                                                                            | Result                                                                                                          |                    |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|--------------------|
| Self-rotation function                                                                                                                               | Two consistent dyad solutions in several resolution ranges                                                      |                    |
| Fc rotation function                                                                                                                                 | Two peaks related by pseudodyad; r.m.s.* = 6.03 and 5.54                                                        |                    |
| (1) Search probe entire Fc                                                                                                                           | same solutions as (1); r.m.s. = 5.66 and 5.44                                                                   |                    |
| (2) CH3 domains                                                                                                                                      | same solutions as (1); r.m.s. = 3.41 and 3.14                                                                   |                    |
| (3) CH2 domains                                                                                                                                      |                                                                                                                 |                    |
| Fab1 rotation function                                                                                                                               | Exceptional peak with constant domains of HYHEL-5; r.m.s. = 6.42                                                |                    |
| (1) Search probes were constant domains of 7 different Fabs                                                                                          | Solution optimal for variable domains of McPC603; r.m.s. = 4.35                                                 |                    |
| (2) Probes were variable domains of 7 different Fabs                                                                                                 | Outstanding solution consistent with (1) and (2); r.m.s. = 7.98                                                 |                    |
| (3) An intact Fab1 constructed according to above results                                                                                            |                                                                                                                 |                    |
| Fab2 rotation function                                                                                                                               | Unambiguous solution for probe with elbow angle $140^\circ$ ; r.m.s. = 6.27                                     |                    |
| Range of Fab models with elbows of $120^\circ$ – $180^\circ$ constructed using XPLOR by altering Fab1 elbow every $5^\circ$                          |                                                                                                                 |                    |
| Translation function                                                                                                                                 | A self-consistent set of solutions from all searches (1)–(4); $cc^\dagger=0.157$ – $0.249$ for 4–8 Å resolution |                    |
| (1) Fc fixed, Fab1 moving                                                                                                                            |                                                                                                                 |                    |
| (2) Fc fixed, Fab2 moving                                                                                                                            |                                                                                                                 |                    |
| (3) Fab1 fixed, Fab2 moving                                                                                                                          |                                                                                                                 |                    |
| (4) Fab1 and Fab2 fixed, Fc moving                                                                                                                   |                                                                                                                 |                    |
| Refinement                                                                                                                                           |                                                                                                                 |                    |
| (1) Rigid body with twelve $\beta$ -barrel domains; 3.5–12 Å                                                                                         | $R=0.386$                                                                                                       | $cc^\dagger=0.529$ |
| (2) Powell minimization and simulated annealing after insertion of correct amino-acid sequence (occupancy of 46 hinge residues set to zero); 3.5–8 Å | $R=0.188$                                                                                                       | $cc^\dagger=0.876$ |
|                                                                                                                                                      | <u>r.m.s. deviations</u>                                                                                        |                    |
|                                                                                                                                                      | Bonds                                                                                                           | 0.017 Å            |
|                                                                                                                                                      | Angles                                                                                                          | $4.038^\circ$      |
|                                                                                                                                                      | Dihedrals                                                                                                       | $28.597^\circ$     |
|                                                                                                                                                      | Impropers                                                                                                       | $0.686^\circ$      |

\* All r.m.s. values are stated for 4–8 Å resolution searches.  $^\dagger$  cc, Correlation coefficient.

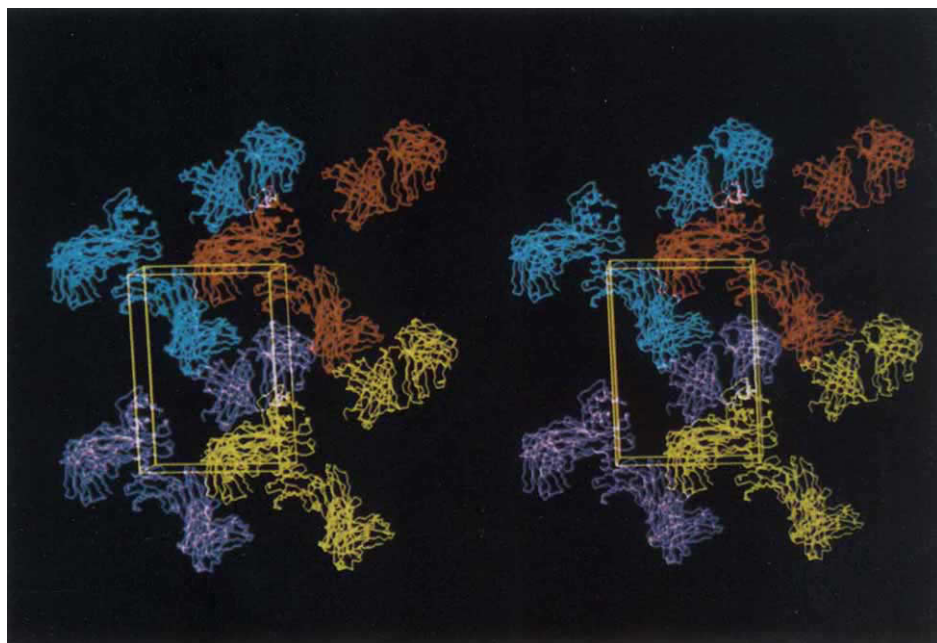
solutions were frequently found even with the probes representing one sixth of the antibody. Rotation function solutions were based on the human Fc fragment<sup>4</sup>, the constant domains of Fab HYHEL-5 (ref. 18), and the variable domains of Fab McPC603 (ref. 19), including the hypervariable regions. Translation searches were performed to determine the relative distances between the three portions of the molecule, the two Fabs and the Fc (Table 1).

The antibody structure was assembled according to essentially independent, but internally consistent, molecular replacement results that ultimately yielded a model fully consistent with the

stereochemistry of an intact antibody. The packing revealed good complementarity of surfaces without interpenetration. Lattice contacts immobilize all segments of the molecule to permit visualization of both Fabs as well as the Fc.

The structure of the antibody is shown in Figs 1 and 2. Its most prominent features are: (1) There is an approximate 2-fold axis relating the heavy chains of the Fc portion of the molecule. The dyad deviates particularly for the CH2 domains; (2) the disposition of the Fc with respect to the Fab portions is quite oblique; (3) the hinge angle between the Fc and Fab1 is approximately  $65^\circ$  and for Fab2 about  $115^\circ$ ; (4) the long axes of the

FIG. 3 Stereo diagram of the packing of four antibodies in the triclinic cell, each of a different colour, showing the intricate network of intermolecular contacts that stabilizes the conformation of the molecule. The Fc segments, which lie more or less along the longest body diagonal, are immobilized by multiple Fab contacts, suggesting why the Fc in these crystals is ordered. The constant domains of an Fab2 of one molecule insert in the elbow region of Fab1 of a different antibody molecule to fix the dispositions of the Fabs. Notable initial exceptions to the otherwise acceptable packing were three hypervariable loops protruding from the variable domains of Fab McPC603. When the correct sequence for the canine lymphoma antibody was examined, it was apparent that the offending residues corresponded to deletions in the latter molecule. Thus, when the correct amino-acid sequence was substituted, virtually all of the packing exceptions were eliminated, as shown in this view.





two Fabs are almost collinear; thus, there is an approximate long axis running through the entire Fab assembly. The angle between the Fabs is  $170 \pm 2^\circ$ , and the Fab axes are offset by  $9^\circ$ ; (5) Fab1 has an elbow angle of  $159^\circ$ , and Fab2 has an elbow of  $143^\circ$ . These elbow angles are near the middle of the range of values observed for other Fabs<sup>1</sup>; (6) the constant domains of Fab1 and Fab2 are related by a near exact dyad axis of symmetry. The variable domains are not so related because of the difference in elbow angles of the Fabs; (7) the dyad of the Fc is at an angle of about  $120^\circ$  with that dyad relating constant Fab domains; (8) the Fc 2-fold axis does not intersect the dyad relating the constant domains of the Fabs, nor does it intersect the approximate long axis of the Fabs; (9) in the crystal, all segments share extensive interfaces which severely restrict the dispositions of neighbours. The contacts, illustrated in Fig. 3, presumably stabilize this particular conformation.

The asymmetric conformation, observed in these crystals of the antitumour antibody, should probably not be considered as a static structure which is maintained in solution. The structure probably represents only one of many possible transient conformations. The unique structure is a product of the intrinsic flexibility of the antibody and the lattice interactions that stabilize this particular distribution of domains. Indeed, electron microscopy<sup>20-23</sup>, fluorescence polarization<sup>24,25</sup> and previous X-ray crystallographic studies<sup>5,26,27</sup> have provided extensive evidence for a wide range of conformations based on segmental flexibility.

The structure we present is instructive in that it illustrates the nature and extent of this structural variability, or dynamic range, which is inherent in the antibody. The Fabs are loosely tethered to a mobile Fc. Each Fab can assume its own elbow angle as its environment or function requires. Somewhat unexpected is the fact that, were the elbow angles the same, the Fabs would be related by an almost exact 2-fold axis that is quite independent of the Fc. It is the disposition of the Fc that disrupts the overall symmetry of the molecule. This is in keeping with the disorder, or multiple orientations of the Fc, observed in the Kol antibody structure<sup>7,8</sup>.

The hinge polypeptides are not really hinges, but rather they are tethers that allow the Fab components to drift from the Fc to bind antigen or potentially allow the Fc to move in such a way to trigger effector functions, such as the activation of complement<sup>25,28</sup>. The connecting polypeptides give the Fabs the freedom to move and twist so as to align hypervariable regions with antigenic sites on large, immobile carriers, in this case tumour cells. The crystal structure visually demonstrates that the antibody is an assembly of units possessing a high degree of flexibility, a molecule suited to the task of scavenging foreign objects or activating a cell lysis system. □

Received 3 August; accepted 9 October 1992.

- Wilson, I. A., Rini, J. M., Fremont, D. H., Fieser, G. G. & Stura, E. A. *Meth. Enzym.* **203**, 153-176 (1991).
- Alzari, P. M., Lascombe, M.-B. & Poljak, R. J. *A. Rev. Immun.* **6**, 555-580 (1988).
- Davies, D. R., Padlan, E. A. & Sheriff, S. A. *Rev. Biochem.* **59**, 439-473 (1990).
- Deisenhofer, J. *Biochemistry* **20**, 2361-2370 (1981).
- Silverton, E. W., Navia, M. A. & Davies, D. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5140-5144 (1977).
- Rajan, S. S. *et al. Mol. Immun.* **20**, 797-799 (1983).
- Colman, P. M., Deisenhofer, J., Huber, R. & Palm, W. *J. molec. Biol.* **100**, 257-282 (1976).
- Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. *J. molec. Biol.* **141**, 369-391 (1980).
- Steplewski, Z., Jeglum, K. A., Rosales, C. & Weintraub, N. *Cancer Immun. Immunother.* **24**, 197-201 (1987).
- Rosales, C., Jeglum, K. A., Obrocka, M. & Steplewski, Z. *Cell. Immun.* **115**, 420-428 (1988).
- Jeglum, K. A. *Proc. Seventh ACVIM Forum, San Diego* (922-925) (Lippincott, Hagerstown, Maryland, 1989).
- Larson, S., Day, J., Greenwood, A., Skaletsky, E. & McPherson, A. *J. molec. Biol.* **222**, 17-19 (1991).
- Fitzgerald, P. M. D. *J. appl. Crystallogr.* **21**, 273-276 (1988).
- Brunker, A. T., Kuriyan, J. & Karplus, M. *Science* **235**, 458-460 (1987).
- Brunker, A. T. *A. Rev. phys. Chem.* **42**, 197-223 (1991).
- Bernstein, F. C. *et al. J. molec. Biol.* **112**, 535-542 (1977).
- Crowther, R. A. in *The Molecular Replacement Method* (ed. Rossmann, M. G.) 173-178 (Gordon and Breach, New York, 1972).
- Sheriff, S. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 8075-8079 (1987).
- Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. *J. molec. Biol.* **190**, 593-604 (1986).
- Wrigley, N. G., Brown, E. B. & Skehel, J. J. *J. molec. Biol.* **169**, 771-774 (1983).
- Roux, K. H. *Eur. J. Immun.* **14**, 459-464 (1984).
- Lamy, J. *et al. Biochemistry* **24**, 5532-5542 (1985).

- Wade, R. H., Taveau, J. C. & Lamy, J. N. *J. molec. Biol.* **206**, 349-356 (1989).
- Schneider, W. P., Wensel, T. G., Stryer, L. & Oi, V. T. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2509-2513 (1988).
- Dangl, J. L. *et al. EMBO J.* **7**, 1989-1994 (1988).
- Huber, R., Deisenhofer, J., Colman, P. M. & Matsushima, M. *Nature* **264**, 415-420 (1976).
- Amzel, L. M. & Poljak, R. J. *A. Rev. Biochem.* **48**, 961-997 (1979).
- Oi, V. T. *et al. Nature* **307**, 136-140 (1984).
- Jones, T. A. in *Computational Crystallography* (ed. Sayre, D.) 307-317 (Oxford University Press, New York, 1982).

ACKNOWLEDGEMENTS. This research was supported by grants from the NIH, from the NSF and from NASA, as well as by the Immunopharmaceuticals Company of San Diego. We thank the San Diego Supercomputer Facility for time on the Cray Y-MP.

## RETRACTION

### Identification by anti-idiotypic antibodies of an intracellular membrane protein that recognizes a mammalian endoplasmic reticulum retention signal

D. Vaux, J. Tooze & S. Fuller

*Nature* **345**, 495-502 (1990)

OUR further characterization of the  $M_r$  72,000 (72K) protein has shown that the data in Fig. 2 of our paper are erroneous and not repeatable. We retract the statement that the 72K protein is an integral membrane protein. The present evidence is consistent with this protein being associated with the intermediate compartment. We also withdraw our speculation concerning its function. □

## GUIDE TO AUTHORS

PLEASE follow these guidelines so that your manuscript may be handled expeditiously.

*Nature* is an international journal covering all the sciences. Contributions should therefore bear in mind those readers who work in other fields and those for whom English is a second language, and write clearly and simply, avoiding unnecessary technical terminology. Space in the journal is limited, making competition for publication severe. Brevity is highly valued. One printed page of *Nature*, without interruptions, contains about 1,300 words.

Manuscripts are selected for publication according to editorial assessment of their suitability and reports from independent referees. They can be sent to London or Washington and should be addressed to the Editor. Manuscripts may be dealt with in either office, depending on the subject matter, and will where necessary be sent between offices by overnight courier. High priority cannot be given to pre-submission enquiries; in urgent cases they can be made in the form of a one-page fax. All manuscripts are acknowledged on receipt but fewer than half are sent for review. Those that are not reviewed are returned as rapidly as possible so that they may be submitted elsewhere without delay. Contributors may suggest reviewers; limited requests for the exclusion of specific reviewers are usually heeded. Manuscripts are usually sent to two or three reviewers, chosen for their expertise rather than their geographical location. Manuscripts accepted for publication are typeset from the London office.

*Nature* requests authors to deposit sequence and crystallographic data in the databases that exist for this purpose, and to mention availability of these data.

Once a manuscript is accepted for publication, contributors will receive proofs in about 4 weeks. *Nature's* staff will edit manuscripts with a view to brevity and clarity, so contributors should check proofs carefully. Manuscripts are generally published 2-3 weeks after receipt of corrected proofs. *Nature* does not exact page charges. Contributors receive a reprint order form with their proofs; reprint orders are processed after the manuscript is published and payment received.

### Categories of paper

**Review articles** survey recent developments in a field. Most are commissioned but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews Coordinator. Length is negotiable in advance.

**Articles** are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text (not including figure legends) or more than six display items and should not occupy more than five pages of *Nature*.

Articles start with a heading of 50-80 words written to advertise their content in general terms, to which editors will pay particular attention. The heading does not usually contain numbers, abbreviations or measurements. The introduction to the study is contained in the first two or three paragraphs of the article, which also briefly summarizes its results and implications. Articles have fewer than 30 references and may contain a few short subheadings.

**Letters** are short reports of outstanding novel findings whose implications are general and important enough to be of interest to those outside the field. Letters should have 1,000 or fewer words of text and four or fewer display items. The first paragraph describes, in not more than 150 words and without the use of abbreviations, the background, rationale and chief conclusions of the study for the particular benefit of non-specialist readers. Letters do not have subheadings and should contain fewer than 30 references.

**Commentary articles** deal with issues in, or arising from, research that are also of interest to readers outside research. Some are commissioned but suggestions can be made to the Commentary Editor in the form of a one-page synopsis.

**News and Views articles** inform non-specialist readers about new scientific advances, sometimes in the form of a conference report. Most are commissioned but proposals can be made in advance to the News and Views Editor.

**Scientific Correspondence** is for discussion of topical scientific matters, including those published in *Nature*, and for miscellaneous contributions. Priority is given to letters of fewer than 500 words.

### Preparation of manuscripts

All manuscripts should be typed, double-spaced, on one side of the paper only. An original and four copies are required, each accompanied by artwork. If photographs are included, five sets of originals are required, for line drawings, one set of originals and four good-quality photocopies are acceptable. Reference lists, figure legends and tables should all be on separate sheets, all of which should be double-spaced and numbered. Three copies of relevant manuscripts in press or submitted for publication elsewhere should be included with each copy of a submitted manuscript, and clearly marked as such. Five copies of revised and resubmitted manuscripts, labelled with their manuscript numbers are required, together with five copies of a letter detailing the changes made.

**Titles** are brief and simple. Active verbs, numerical values, abbreviations and punctuation are to be avoided. Titles should contain one or two key words for indexing purposes.

**Artwork** should be marked individually and clearly with the author's name and, when known, the manuscript number. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are to be avoided and are permitted only if the parts are closely related, either experimentally or logically. Unlettered originals of photographs should be provided. Suggestions for cover illustrations, with captions and labelled with the manuscript number, are welcome. Original artwork is returned when a manuscript cannot be published.

**Protein/nucleotide sequences** should ideally be in the three-letter and not the single-letter code for amino acids. One column width of *Nature* can accommodate 20 amino acids or 60 base pairs.

**Colour artwork.** A charge of £500 per page is made as a contribution towards the cost of reproducing colour figures. Inability to pay these costs will not prevent publication of essential colour figures if the circumstances are explained. Proofs of colour artwork may be sent to contributors under separate cover from their galley proofs.

**Figure legends** should not exceed 300 words and ideally should be shorter. The figure is described first, then, briefly, the method. Reference to a method published elsewhere is preferable to a full description. Methods are not described in the text.

**References** are numbered sequentially as they appear in the text, followed by those in tables and finally by those in figure legends. Only papers published or in the press are numbered and included in the reference list. All other forms of reference should be cited in the text as a personal communication, manuscript submitted or in preparation. Text is not included in reference lists. References are abbreviated according to the *World List of Scientific Periodicals* (Butterworths, London, 1963-65). The first and last page numbers are included; reference to books should include publisher, place and date.

**Abbreviations, symbols, units and Greek letters** should be identified the first time they are used. Acronyms should be avoided wherever possible and, if used, defined. Footnotes are not used in the text.

**Acknowledgements** are brief and appear after the reference list; grant and contribution numbers are not allowed.

**Supplementary information** is material relevant to Articles or Letters which cannot, for lack of space, be published in full, but which is available from *Nature* on request.

**Submission.** Manuscripts can be sent to the Editor at 4 Little Essex Street, London WC2R 3LF, UK or at 1234 National Press Building, Washington, DC 20005, USA. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value \$5' to prevent the imposition of import duty and value-added tax.

# Global communicator

Ed Regis

**Odyssey: The Authorised Biography of Arthur C. Clarke.** By Neil McAleer. Contemporary Books/Gollancz: 1992. Pp. 430. \$25, £16.99.

**How the World Was One: Beyond the Global Village.** By Arthur C. Clarke. Bantam/Gollancz: 1992. Pp. 289. \$22.50, £15.99.

NEIL McAleer gives us the bad news right away, in his preface to *Odyssey: The Authorised Biography of Arthur C. Clarke*: "This writer never probed too deeply into Clarke's private affairs. I no more wanted to describe intimate details of a living man's private life (assuming the information was available) than he or his contemporaries wanted to read about them. If this is the compromise of an 'authorised' life, so be it."

What this means is that we're not going to hear about Clarke's sex life. Nonetheless, despite this disclaimer, McAleer presents a full account of Clarke's whirlwind, two-week engagement ("Spontaneous combustion", he said; "Electric", she said) and fleeting, six-month marriage to a young American divorcée, as well as his later musings about sexuality in a *Playboy* interview: "I think Freud said something to the effect that we're all polymorphously perverse, you know. And of course we are", Clarke says, adding (when asked) that he has had bisexual experiences: "Who hasn't? Good God! If anyone had ever told me that he hadn't, I'd have told him he was lying."

Arthur Charles Clarke, one gathers, is a man of unequivocal and strong opinions. A farmer's son from Somerset, he developed a taste for exploration early on. "The north coast of Somerset offers vistas of the Atlantic Ocean, which created the illusion of infinite space", McAleer says, "and there was the regular excitement of ships leaving port and beginning voyages to distant lands."

Space, ocean, distant lands — these were the main ingredients of Clarke's life and work. He became a science fiction fan at the age of eleven, when he read an issue of *Amazing Stories* at a

friend's house, and was soon writing science fiction himself, an enterprise that McAleer improbably construes as Clarke's "search for his missing father", who died when the boy was thirteen.

Clarke was an early member of the British Interplanetary Society, and built homemade rockets ("Why get excited about anything that doesn't go straight

ped out because he found it "extremely dull", and took a job instead as assistant editor of *Physics Abstracts*, work that gave him a far better education: "I probably had a bird's-eye view of research in physics unmatched by anyone else on Earth during this period."

One of his first nonfiction books, *The Exploration of Space*, was a Book-of-the-Month Club selection in the United States, for which he got a big advance and great reviews, and from that moment on his career was made. Ever since, Clarke has been leaving drab terra firma in one form or another, whether through science fiction, nonfiction writing about space travel, or scuba diving. "Mr Clarke is not worldly; he is other-worldly", says an acquaintance.

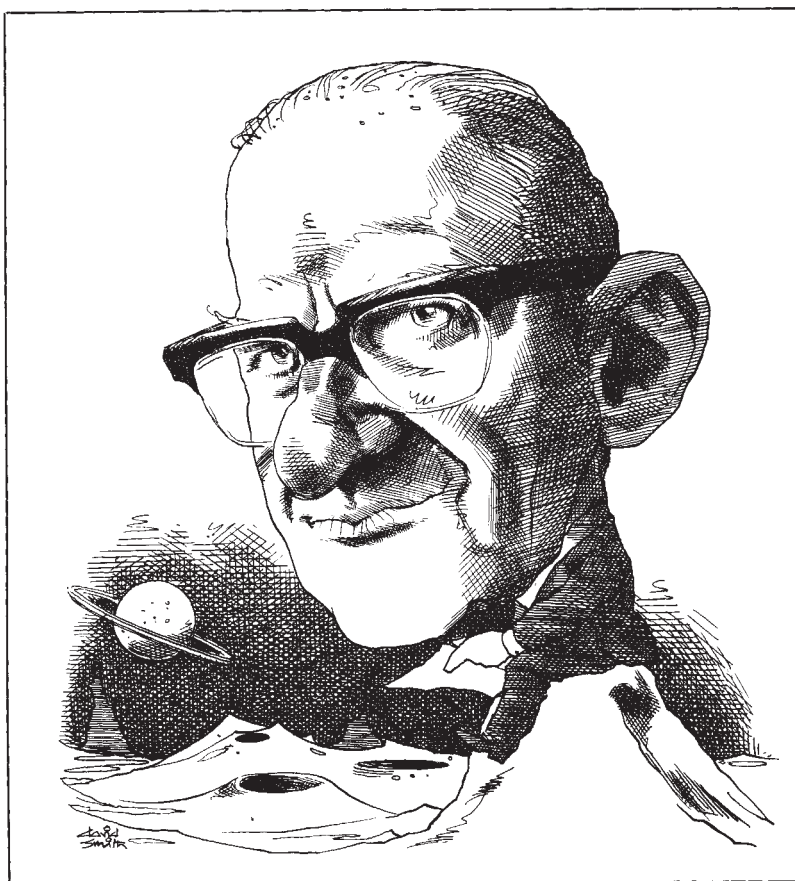
Throughout, Clarke emerges as a man of warmth, wit and good cheer, with a vast tolerance for human foibles and ironies. He cried when he watched Apollo II leave for the Moon, but laughed when he noticed what one of the television technicians was doing while Neil Armstrong and Buzz Aldrin walked on the lunar surface: "He was reading the racing reports", Clarke recalled.

He had tolerance as well for the foibles of his own body. A bump on the head in 1962 led to temporary paralysis, ultimately diagnosed as post-polio syndrome, an ailment from which he never fully recovered. Nevertheless, his innate happiness and optimism never left him, and it was after this episode that J. B. Priestley described him as "the happiest writer I have ever met". Indeed, it is hard to find a picture in this well-illustrated book in

which Clarke is not smiling, "a beaming shambles" according to a friend.

*Odyssey* is an anecdote-filled, compulsively readable story of Clarke's own voyage through the twentieth century and beyond. It tells of his inventions and prophecies and, toughest of all, manages to rekindle the sense of wonder and awe from the early days of spaceflight, when the words "Live from the Moon" were broadcast on television for the first time.

It also tells the story of *2001: A Space Odyssey* and its famously unintelligible ending. He and director Stanley Kubrick couldn't decide how to close the movie,



up?"), telescopes and — "the thing I'm proudest of" — a device that, by means of a carbon microphone, bicycle lamp and photoelectric cell, transmitted the human voice across short distances. Communications technology proved to be Clarke's other main passion: he was, and remains, the world's foremost telecommunications evangelist.

Clarke attended King's College London where he got a Bachelor of Science degree in physics, pure mathematics and applied mathematics, and then enrolled at University College London for a postgraduate degree in astronomy. He drop-



but finally Clarke came up with an ending. There is no indication that either of them ever knew what it meant, or cared very much one way or the other. "I don't like to talk about 2001", said Kubrick; "I don't pretend we have the answers", said Clarke.

Clarke later wrote a succession of sequels and other 'last books', and one might suppose that *How the World Was One: Beyond the Global Village* is one of them. But in truth it's mostly an early book repackaged. Although it is not mentioned anywhere on the dust jacket, *How the World Was One* is largely a revision of Clarke's book *Voice Across the Sea*, first published in 1958. The rest consists of addresses given on various occasions, three previously published short stories and, for good measure, the two scientific papers in which Clarke proposed the idea of communications satellites. Not that there is anything wrong with an author's recycling of old materials — just so long as the prospective buyer is informed of the fact beforehand. Notwithstanding that there are fewer than 50 pages of new material here, Clarke has described *How the World Was One* as "My longest book . . . . It may also be one of my most important, since it's covering the whole history of telecommunications, from the invention of the telegraph, up to the fibre optic revolution."

The whole history is indeed there, and the first part of the story, dealing with the laying of the transatlantic cable, makes exciting reading. "Wiring the abyss", as Clarke puts it, was one of those impractical utopian fantasies that everyone knew could never be realized. Often enough it seemed that way even to the participants, chief among whom were Cyrus W. Field, who financed the project, and Isambard Kingdom Brunel, who built the ship, the *Great Eastern*, that laid the cable. Much of their tale is cliffhanging stuff. The first attempt to lay the cable, in 1857, met with failure when the cable snapped and was lost after 335 miles had been paid out. The second try, in 1858, was successful for a three-week period, after which the line went dead. The third try, in 1865, went well until the cable parted in the middle of the ocean and dropped to the bottom. Incredibly, the broken end was snagged by a grappling hook and was almost retrieved, but at the surface the rope broke and the whole mess slipped back to the ocean floor, two miles below, where it was not found again for several months.

The rest of the book, unfortunately, is not nearly so compelling, describing how a gaggle of communications satellites were lofted into orbit on a series of anonymous rockets. Instantaneous global communication is now something

that we take so much for granted that the story has lost its lustre.

One may wonder, then, why Clarke is still so entranced by it. His answer, made clear in the text time and again, is his belief that instantaneous global communication is the royal road to peace among men. Satellite talk will make the world 'One', turning us into the Big Global Village, the Planetary Happy Family. This, though, is a delusion. Whereas Clarke's hope is that communications satellites will promote "understanding", the fact is that they will promote whatever messages are bounced off them, and it is just as easy to send orders to assassinate a leader or wipe out a village as it is to bring tidings of universal peace and brotherhood. Nowhere was this better illustrated than during the Gulf War, instantaneous live coverage of which Clarke portrays as showing us "the true horrors of modern

mechanised war". On the contrary, what it showed us were absolutely thrilling images of 'smart' bombs homing in on and destroying office buildings — images that were in large part responsible for making this one of the most popular wars in American history.

As for the idea that One World means unity and peace, sufficient disproof is offered by the existence of civil wars, which have been some of the bloodiest ever fought on the planet. Anyone who thinks that One World (or One Nation or One Anything Else) will bring us peace and harmony should spend a few silent moments contemplating the gravestones at Gettysburg Battlefield, where during the American Civil War some 51,000 casualties were sustained in three days. □

*Ed Regis is at Western Maryland College, Westminster, Maryland 21157, USA.*

## Physicist in the shadows

*Hermann Bondi*

**The Recollections of Eugene Wigner as Told to Andrew Szanton.** By Andrew Szanton. Plenum: 1992. Pp. 335. \$24.50.

HAVING as its subject a man who is not only an outstanding physicist, but also most unusual and indeed idiosyncratic in his attitudes and outlook, this volume should have been superb and gripping, yet in fact it does not generally rise to these heights. Perhaps Andrew Szanton was too self-effacing and too little of the probing professional interviewer who would have coaxed more out of his subject. But then, if he had taken such an approach, Wigner might not have talked to him at all. We should therefore all be grateful to Szanton, because this book is likely to remain the most we will ever know about a fascinating person. It deserves to be read by all with even a modest interest in the personalities who shaped modern quantum physics.

Despite my little gripe, much of the book is absorbing. The early chapters about Wigner's home life and upbringing are thorough and accordingly riveting. Especially moving are his feelings about his school, his teachers in general and his mathematics teacher in particular. Nobody who reads these chapters can have any doubts that an outstanding teacher can deeply influence and foster the abilities of able pupils who without such a stimulus might never rise to great heights of achievement, whatever their innate abilities. How much talent is undiscovered, unstimulated, unfostered in our world? This is a worrying, sobering

thought generated by these chapters.

The life of a Jewish professional family in Hungary in those far-off days comes over well, but I was perturbed by confusion in the text about what was Austro-Hungarian and what was Hungarian in the days of the Dual Monarchy, a matter that is likely to disturb other readers less than me. But the brief paragraph on how such a civilized and humane family could come to terms with living under the blood-stained Horthy regime is superb. I recall only too well how dreadful it looked from Austria.

Although the description of the chain of circumstances that made Wigner a physicist against the wishes of his father makes excellent reading, the account of the rest of these years of apprenticeship is too brief and too scrappy. One hungers for more about the giants of the time and their interaction, but one is given only brief vignettes. Illuminating flashes occur and are a delight. To hear that Dirac's work was found difficult because he wrote in English gives one a healthy shock. For few recall now how dominant German was as the language of science for a period that Hitler brought to a sharp end. It would make a good study to analyse how the language of Newton, Faraday and Maxwell came to be quite secondary to German for some decades before it acquired its present dominance. But again, although there are splendid shafts of light describing Dirac's unique attitude to conversation, one would have hoped to hear more about an extraordinary man who was not only Wigner's friend and colleague but also his brother-in-law. But

the biggest problems in these sections about the pre-war period are due to Wigner's obtrusive modesty. The overly shy description of what he did in those years gives no clue that it was work of Nobel-prize calibre.

The book really comes alive again with the chapters dealing with Wigner's involvement in the Manhattan Project. These are quite excellent. They will be particularly appreciated by all those unacquainted with the workings of bureaucracy, especially if they can sense how much it can eventually achieve. The remarks about General Groves are splendid, as are the sketches of Fermi and other scientists. Wigner is a little less self-effacing about his own contribution here, but even so I would have liked to learn more about it. On the other hand, Wigner's ruminations about the work that was going on in Germany on the atomic bomb are generous and thought-provoking, as are his revealing comments on Oppenheimer.

The post-war period is described in an illuminating way. Wigner could come to terms with a bureaucracy driven by the pressures of war, but not with one in peacetime. The return to Princeton was a success, which enabled him to do much external work in which he evidently found some satisfaction and some frustration. He is at his most engagingly frank when talking about the Nobel prize and the pleasure it gave him.

Perhaps the most revealing paragraphs relate to his fellow Hungarians Szilard, von Neumann and Teller. His tortured love-hate relationship with Szilard evidently troubled him greatly, and the honesty of his remarks is transparent. His admiration for von Neumann and his grief for his early death come through very well, but inevitably it is his comments on that very controversial character, Teller, that hold most interest. Wigner's loyalty to Teller is total, but he manages to combine it with his admiration for Oppenheimer's leadership at Los Alamos in a frank and indeed engaging manner. What informed Wigner's Hungarian group was a permanent deep hatred of communism (perhaps because of the experience of Bela Kun?), never ameliorated by the slim hope that so many of us entertained for so many years that something less obnoxious, even if still ostensibly communist, might eventually emerge in Eastern Europe.

This brief description cannot possibly do justice to an important and very worthwhile book. My disappointments, which I have expressed freely, should not discourage anybody from reading it. In the interests of science I wish this book every success. □

Sir Hermann Bondi is at Churchill College, Cambridge CD3 0DS, UK.

## From thought to expression

S. S. Schweber

**Genius: Richard Feynman and Modern Physics.** By James Gleick. *Pantheon/Little, Brown*: 1992. Pp. 532. \$27.50, £18.99.

SCIENTIFIC communities usually seek to convey the importance and the history of their discipline through the lives of their outstanding practitioners — individuals whose very creativity renders them unlikely to be the best representatives of their worlds. Physicists have principally chosen theorists as their heroic figures (Newton, Maxwell, Einstein, Planck, Bohr, Dirac, Pauli) or experimenters who left their mark in both experimental



Talking it through — Feynman chats with a student as Murray Gell-Mann looks on.

tion and theory (J. J. Thomson, Rutherford, Fermi), thereby emphasizing the ties of the discipline to the tradition of natural philosophy. The cult idols are the reification of the aspirations and myths of the community and the embodiments of its cherished values. (And just as what is not said can be as revealing as what is, those left out of the hall of fame disclose much about the community.)

Einstein and Bohr, and to a lesser degree Planck and Rutherford, are the heroes from the period before the Second World War. They have come to stand for the advances ushered in by relativity and quantum mechanics. Dirac embodies the 'genius' of the quantum-mechanical revolution and is the prophet who pointed the way in the synthesis of relativity and quantum mechanics. And Landau and Feynman have become the heroes of the post-quantum-mechanical era. Both were enormously creative and innovative and made lasting

conceptual contributions (a necessary condition for admission to the pantheon). Both encompassed all of physics, and could therefore be taken as representing the unity of the enterprise, resisting the fragmentation of the discipline and the attendant specialization. Although at various times they set and dominated the intellectual agenda of the community, they were often also the main challengers of the ruling orthodoxies. They were thus simultaneously insiders and outsiders. Similarly, both were unusual in their social behaviour — and the community was willing to condone in them conduct that would not be acceptable in society at large. And after the Second World War, during an epoch shaped by the Cold War, polarized by concerns about national security and dominated by military-industrial complexes, both men consciously distanced themselves from the conflict. They could therefore be seen as 'pure scientists' transcending national rivalries.

There is so far no satisfactory intellectual and psychological portrait of Landau or an adequate account of his life. But with James Gleick's *Genius* we now have a superb biography of Feynman. The book is a moving, beautifully written, literate and perceptive account of Feynman's life, which Gleick sets sensitively in the context of the physics community, of the larger culture and of the times.

We get a feel for life in the Jewish community of Far Rockaway during the

1920s. We learn what it was like to be an undergraduate at the Massachusetts Institute of Technology during the 1930s. We become privy to the extensive antisemitism in the élite American universities at that time; to life at Los Alamos during the Second World War; to the politics of universities after the war; to the dynamics of selecting and receiving Nobel prizes; to the world of NASA stripped of its public-relations gloss, and the behind-the-scenes activities of the presidential commission investigating the Challenger disaster.

As readers of Gleick's earlier book *Chaos* know, this author has a gift for capturing the atmosphere of a community and for providing sharply focused, perceptive sketches of its principal characters. We get to know Schwinger, Bethe, Weisskopf, Dyson, Gell-Mann and some of the other leading theorists, and, while doing so, we learn a great deal about the sociology of the theoretical physics community of the time and



the competitive dynamics that animated it. The book also contains lucid accounts of many of the advances in theoretical physics since the middle of this century — Feynman diagrams, renormalization theory, the explanation of the superfluidity of liquid helium, partons and quarks, the standard model — advances to which Feynman made important, fundamental contributions. Gleick's accomplishment in this area is uncommon because he is not a scientist by training, yet his presentation of the science is accurate and readily accessible to the general reader.

Gleick never met Feynman. His knowledge of his subject stems from interviews with members of Feynman's family, with colleagues and former students, and with friends dating back to Feynman's highschool and college days. He read and listened to taped interviews with Feynman that various historians had made earlier, in particular the extensive and remarkable interviews that Charles Weiner recorded in the late 1960s and mid-1970s. He studied videotapes of Feynman's lectures and television programmes, and pored over some 25 cartons of papers, notes and correspondence that Feynman had deposited in the archives of the California Institute of Technology. It is clear that Feynman was unusually gifted, highly original and inventive. But the vivid portrait that Gleick draws would not have been possible without access to the personal letters that Feynman had kept in his house. Gweneth, Feynman's wife, allowed Gleick to read the correspondence between Feynman and his parents, between Feynman and his first wife, Arline, and letters to Feynman from lovers and personal friends.

Feynman's relationship with Arline — his first girlfriend, his first and perhaps his only true love, and his first wife — shaped his subsequent life. Arline was diagnosed as having lymphatic tuberculosis while Feynman was at MIT. Despite the strenuous objections of his parents, the two got married before Feynman went to Los Alamos early in 1943. Throughout his stay there, Arline was in a hospital in Albuquerque and she died shortly before Trinity in 1945. Feynman never recovered from the loss, seemingly arriving at the conviction that he would never again be fulfilled in love. Two years after Arline died, Feynman — the supreme rationalist — wrote her a heart-rending love letter. His subsequent interactions with women were always affected by the scars incurred by the loss — until his marriage to Gweneth, the emptiness and yearning were filled by one-night stands and tempestuous, destructive love affairs. Similarly, in physics he came to accept human limitations: unification and, in particular, a theory of



Genius born or made?

everything were, in his view, fantasies with which the community deluded itself. Physics consisted of a set of algorithms that answered with a high degree of certainty questions of the form: 'If I do this, what will happen?' Feynman searched for such algorithms with extraordinary courage and unshakeable integrity.

Gleick is captivated by Feynman's "genius". (Who wouldn't be?) At times he seems obsessed by the notion of genius and by the search for what constitutes its possession. He fails to find the holy grail, but does make clear that to understand Feynman one has to apprehend his peculiar connection to his social world and his distinctive intellectual relationship to the scientific community in general and the physics community in particular. In intellectual matters Feynman was able to straddle the gulf between self and community. He could adhere to many of the tenets, assumptions, forms of thought and styles of reasoning that characterized the theoretical physics of his day while also transcending their limitations. One aspect of Feynman's genius was that he could make clear what was obscure to most of his contemporaries. His doctoral dissertation and his 1947 *Reviews of Modern Physics* article that presented his path-integral formulation of nonrelativistic quantum mechanics helped to clarify in a striking manner the assumptions that underlie the usual quantum-mechanical description of the dynamics of microscopic entities. And he did this in the very act of transcending the usual formulation with a startling innovation. It may well be that his reformulation of quantum mechanics and his 'integral over paths' will turn out to be his most profound and enduring contributions. They have deepened considerably our understanding of quantum mechanics and have greatly extended the systems that can be quantized. And judging from the work of Michael Atiyah and Ed

Witten, Feynman's path integral has already substantially enriched mathematics and provided new insights into spaces of infinite dimensions.

Early on, Feynman also learned to walk the tightrope between his own psychological needs and the requirements of belonging to a community. He came to accept and appreciate the fact that the act of creation was for him also an act of consummate isolation.

The creative act depends on private visions and solitary constructions and always draws on the legacy and the resources of the community, be it in the arts, literature, technology or the sciences. We call people geniuses when their ability to synthesize these communal resources overwhelms us; and if the synthesis happens to result in a startling outcome, as in the case of Feynman, we are amazed and awe-struck. But by creating a category of people we label as geniuses, we are on a slippery slope that may lead us to believe that there is an innate quality attached to the attribution, when instead we should focus on its social and cultural dimensions. Rather than *Genius*, I would have called this impressive book *The Remarkable Life and Science of Richard Feynman*. □

S. S. Schweber is in the Department of Physics, Brandeis University, Waltham, Massachusetts 02254, USA.

## Mathematics in the mind

Michael Berry

**Pi in the Sky: Counting, Thinking, and Being.** By John D. Barrow. Oxford University Press: 1992. Pp. 317. £14.95, \$25.

MY son, a musical beginner, keeps getting confused about intervals: asked to play a fourth, he delivers a fifth. I suppose this is a widespread confusion, whose origin might lie in the absence of zero from musical notation. Intervals (n) therefore obey not the addition rule of ordinary arithmetic but  $(n)+(m)=(n+m-1)$ : a second plus a second equals a third. Realizing this, we suddenly see how deeply we have interiorized the conventions of elementary mathematics. One of John Barrow's aims is to trace the origins of these counting conventions in human (and animal) societies. This is part of a far more ambitious programme: understanding the meaning of mathematics and the role it plays in science.

By profession, Barrow is an astronomer and physicist, and the mainspring NATURE · VOL 360 · 26 NOVEMBER 1992

of this work is what Wigner called "the unreasonable effectiveness of mathematics". Wigner was referring to the mysterious phenomenon in which areas of pure mathematics, originally constructed without regard to application, are suddenly discovered to be exactly what is required to describe the structure of the physical world. Thus, Riemann's general formulation of the geometry of curved spaces was essential to Einstein's understanding of gravity; Heisenberg found that the symbolic arrays which in quantum mechanics represent observable quantities were the matrices that had been invented decades earlier; and now recondite aspects of the distribution of prime numbers might well provide the link between quantum mechanics and newtonian chaos.

Such connections raise many questions. Is mathematical truth invented by mathematicians, or does it already exist in the world, to be discovered when our minds become sophisticated enough? If discovered, where is it beforehand? What is its relation to the matter whose behaviour it describes so well? Is there any inapplicable mathematics?

Barrow does not answer these questions, but gives a careful and perceptive account of their background and the philosophies they have stimulated. He starts, appropriately enough, with an anthropological and historical analysis of counting and calculation, focusing on the tricky question of whether such skills are innate, and would inevitably develop in any human society, or whether they arose 'accidentally' in one (or several) societies, and diffused to the others. The latter is, he thinks, more plausible. Central here are the inventions (discoveries?) of place values and of zero, by the Babylonians and Hindus 5,000 years ago, leading via the mediaeval Arabs to the decimal system we use today.

Because mathematics is the most precise embodiment of systematic thought, it was natural to try to prove that it has a solid foundation in logic and is perfectly consistent. The story of these attempts has often been told. How Frege, Russell and Whitehead tried to 'derive' mathematics from logic almost a century ago, and how this attempt was complicated by the irritating paradoxes of self-referential sets ('If the barber shaves everyone who does not shave himself, who shaves the barber?'). How Hilbert took up the challenge by trying to prove the consistency of mathematics from within, by formalizing its symbols and deductive steps. How "all the noonday brightness of this confident picture of the formalists' little mathematical world was suddenly extinguished" by Gödel's proof in 1931 that the set-theory paradoxes make it impossible for a sufficiently complicated system to be proved consistent

from within. These ideas are central to modern notions of randomness as the inability to compress information, and may have implications for our attempts (in my view doomed) to find a compact encoding of the physical universe as a 'theory of everything'. Barrow's account of these matters is lucid and engaging.

After pointing out that "formalism is lacking in two crucial respects" (it does not explain the usefulness of mathematics and its relation to the minds of mathematicians), Barrow turns to inventionism. This "amounts to the claim that mathematics is a branch of . . . psychology". It makes "mathematical truth dependent upon time and history", and "one cannot help but feel that humanity is not really clever enough to have 'invented' mathematics".

A chapter is devoted to Brouwer's programme of intuitionism, where the natural numbers are regarded as unarguably 'given', and the aim is to build the rest of mathematics "by step-by-step deductions using a finite number of steps". This brought him into collision with Hilbert, who believed that such a philosophy, which disallowed infinite processes such as arguing by *reductio ad absurdum*, would fatally impoverish and weaken mathematics. Hilbert's attempt to enforce political correctness and to expel Brouwer from the editorial board of *Mathematische Annalen* provoked an absurd and bitter controversy that Einstein called the "war of the frogs and mice".

Finally, Barrow explores the Platonic position that mathematical abstractions exist "in a realm of non-spatial, non-mental, timeless entities". He concludes, albeit somewhat uneasily: "Our ability to create and apprehend mathematical structures in the world is merely a consequence of our own oneness with the world".

I admit to finding some of Barrow's arguments hard to follow not because of their content but because of his habit of using very long sentences unadorned by punctuation whose verbs are hard to find and whose meanings therefore hard to unravel. Worse, some sentences are incomplete, and there are many spelling mistakes. Quotations abound. Some are witty and apposite, but why propagate Spiro Agnew's abysmal "An intellectual is a man who doesn't know how to park a bike"?

These are, however, minor criticisms, and I warmly recommend Barrow's brave attempt to gather up the many loose threads of this elusive subject — a subject so central to our scientific culture — and to grasp the whole of it. □

Michael Berry is in the Department of Physics, University of Bristol, Bristol BS8 1TL, UK.

## Einstein as lover

Joseph Schwartz

**Albert Einstein and Mileva Marić: The Love Letters.** Edited and with an introduction by Jürgen Renn and Robert Schulmann. Translated by Shawn Smith. Princeton University Press: 1992. Pp. 107. \$14.95, £12.50.

THIS elegantly published volume of letters between the young Einstein and the young Marić is a spin-off from the first two volumes of a planned 35 volumes containing some 43,000 documents lying in the Einstein archive. A lovely introduction by Jürgen Renn and Robert Schulmann, coeditors of the project, draws our attention to the unique personality of Marić and her central contribution to the Einstein success story. The meticulous scholarship of the notes is wonderful, particularly the inclusion of the dates of virtually all the characters in this first act of the Einstein drama. And the letters themselves are a treat, a window into the early development of the man who became the most celebrated scientist in history. But what, when all is said and done, does this correspondence tell us?

The Einstein we see here is bubbly optimistic, reassuring, high-spirited, confident about life. For the first time we have an Einstein with sexuality: "Oh my! That Johnnie boy!/So crazy with desire/ While thinking of his Dollie/His pillow catches fire" (letters 19); "How beautiful it was the last time you let me press your dear little person against me in that most natural way" (letter 33). Albert is happy in his sexual relationship with Marić and the letters show it.

There is a not entirely happy story here, however, about two lovers, one who thrives, the other who gets increasingly submerged by life. We meet them both as students of physics. She, a late entrant from the distant provinces of undeveloped Serbia, is three-and-a-half years his senior. He is youthful, exuberant. No obstacle is too great. She, while available for emotional and sexual involvement, is unhappy, feeling that her provincial background has irreversibly limited her chance in physics. While Einstein is absorbing with great fascination the nuts and bolts of doing physics, Marić is distant, observing wistfully the spectacle of her university lecturers: "human beings are so clever and have accomplished so much as I have observed once again here in the case of the Heidelberg professors" (letter 1).

As we journey with these lovers over a



period of five years, we see Marić becoming more unsure of herself, unsure of her place in Einstein's life, pregnant, carrying their illegitimate child on her own while Einstein is employed elsewhere, giving birth without him being there and painfully giving up the baby Liserl for adoption. The letters end with Marić pregnant again, Einstein working at his job in the Swiss patent office, her worried that he will be angry about her being pregnant, he reassuring her: "I'm not the least bit angry that poor Dollie is hatching a new chick. I'm happy about it and had already given some thought to whether I shouldn't see to it that you get a new Liserl. After all you shouldn't be denied that which is the right of all women" (letter 54).

One gets the unmistakable impression that Einstein grew stronger through his relationship with Marić while she increasingly felt weakened and despairing. There is a deep melancholy in this picture of Einstein's life before he became a star.

Like most collections of letters of

leading public figures, these are well worth reading. But we must ask: what are we doing delving into Einstein's love life? Aren't we like Madonna fans, fascinated by our star's stardom? Do these letters tell us what we really want to know?

To understand Einstein we need to understand stardom. We need not the Einstein papers but the newspapers. What has been the role of the media in creating the Einstein legend? What have we been responding to in this myth of the man without socks, the mysterious icon of pure thought? These letters show that, like the luminiferous ether, the Einstein we seek does not exist except in our minds. The mystery of Einstein is the way we have made him incomprehensible. Einstein's achievement, a masterpiece of human understanding, has been turned on its head to become a symbol for the impossible to understand. How did this happen? □

*Joseph Schwartz is at 2 Lancaster Drive, London NW2 4HA, UK.*

to a guest who is entertaining, slightly relentless and — let's face it — a mite opinionated. The superiority of science is ruthlessly asserted. "Scientific knowledge is special and privileged — in the sense that it provides our best understanding of the world." One cannot deny that science provides our best understanding of some aspects of the world, but its success is purchased by the limitation of its ambition. Essentially it is concerned only with certain kinds of impersonal, largely repeatable experience. But a painting is much more than a collection of specks of paint of known chemical composition, and there is a great deal more to human experience than science is able to tackle. Wolpert makes the astonishing mistake of equating the method of investigation with the actual nature of reality. He says, "Any philosophy that is at its core holistic must tend to be anti-science, because it precludes studying parts of a system separately".

Even within science, it is absurd to adopt such a reductionist stance. What if there are holistic laws of nature, such as organizing principles working in the direction of increasing complexity? They will have to be sought through new methodologies, but our concern as scientists must be to respond adequately to the way the physical world actually is. In fact, twentieth-century physical science has seen the death of mere mechanism and the discovery of an interconnectedness (nonlocality) in the fabric of the world.

Wolpert is at his worst when he speaks of religion. There is an assertive dismissiveness ("religious belief is incompatible with science"), derived from a caricature picture ("religion is based on unquestioning certainties"). His treatment of theological thought is as crude an abuse as is the creationists' misuse of scientific thought. Wolpert acknowledges with Tolstoy that "science does not tell us how to live". His answer to moral issues seems to be ultimately the social endorsement of the *vox populi*. Yet his sensitive discussion of the mistakes of the eugenics movement shows that he would not have accepted such policies even if they were endorsed by society (as they were in Nazi Germany). He should think a bit more about what is the source of our intuition of the value of human individuals.

Beneath the civilized discourse of this entertaining book there is a note of unconscious arrogance. The science is tinged with scientism ('science is all') in a way that fuels the fires stoked by the likes of Brian Appleyard or Mary Midgley. □

*John Polkinghorne is President of Queens' College, Cambridge CB3 9ET, UK.*

## Scientism disguised?

*John Polkinghorne*

**The Unnatural Nature of Science: Why Science Does Not Make (Common) Sense.** By Lewis Wolpert. *Faber and Faber*: 1992. Pp. 191. £14.99, \$22.95.

THE unnaturalness of science is held to lie both in the superior clarity of its thought over everyday notions (exemplified, for instance, by common-sense misapprehensions about probabilistic reasoning) and in the counterintuitive character of regimes far from common experience (such as the quantum world). This unnaturalness is to be commended, and Lewis Wolpert's book is a kind of hymn of praise from one of science's practitioners. He rightly distinguishes science from technology, characterizing the scientific aim as the understanding of the world, not its manipulation. It all started in Greece, but Wolpert acknowledges that an essential development occurred in seventeenth-century Europe with the turn to empirical investigation. He is sufficiently candid to recognize that there was religious encouragement to regard the world as rationally structured, but he fails to notice that the idea of the Creator's freedom of action, enshrined in the Judaeo-Christian-Islamic tradition, implied that one had actually to look to see what order He had chosen to create.

Wolpert is good about scientific creativity, recognizing that it requires a

great deal of initial hard work and commitment to the problem before the Damascus-road moment of inspirational insight is able to come. The role of luck is largely rejected; in Pasteur's words, the happy accident is only appreciated by the 'prepared mind'. There is an honest acknowledgement of the scientist's ambition to acquire reputation: "the admiration of one's peers is one of the major rewards of science". Yet, it seems to be that such recognition is fleeting. "Compared to the creative arts, science is ultimately an anonymous enterprise".

The attainments of science have been subject to much reassessment by the twentieth-century masters of suspicion. Philosophers have frequently denied that science can tell us what the physical world is actually like. In response to "relativism rampant", Wolpert finds that not only cheerfulness but common-sense keeps breaking in. He confesses himself to be a common-sense realist in such matters. I am entirely sympathetic with the realist response, but it has to be more critical and nuanced than the discussion in this book. A remark such as "It is important — indeed essential — to separate evidence from theory", totally fails to recognize the theory-laden character of observations.

Wolpert's style is to write at a cracking pace, interlarding the discourse with plenty of anecdotes. Reading his book is like finding oneself sitting at dinner next

379



other animals. Assigning a privileged position to the brain is not 'scientific', but is grounded in the mind-body split that dominates our culture. And there is no scientific reason why killing animals is permissible, when killing humans is not. The distinction does not originate in biology or evolution, but is part of the Judaeo-Christian tradition in which God gave humans dominion over the rest of creation. Buddhist and other traditions do not accept it. Finally, the authors finesse a question that is central to the abortion debate when they mention in passing, and without comment, that taking a human life is murder only when it is done "without the sanction of the state" (page 12). This caveat raises the question why the state should forbid abortion when it sanctions killing people convicted of certain crimes, glorifies killing in wartime and permits it in self-defence. Indeed, why isn't terminating

an unwanted pregnancy a form of self-defence?

Because of their exclusive focus on "the fetus", the authors ignore the fact that, whatever the law says, as long as there are unwanted pregnancies there will be abortions. The important question is whether abortions are legal and safe, or illegal and damage or kill women. Where abortion is legal, it involves fewer risks for women than does a full-term birth. An "objective" book about abortion should consider the effects that gestation, birth and abortion have on women. Scientific information about these topics bears on political decisions about abortion, but this book omits such facts of life. □

*Ruth Hubbard is in the Department of Cellular and Developmental Biology, Harvard University, Cambridge, Massachusetts 02138, USA.*

## Responsible genetics

Dorothy Nelkin

**Gene Mapping: Using Law and Ethics as Guides.** Edited by George J. Annas and Sherman Elias. Oxford University Press: 1992. Pp. 291. \$39.95.

MORE than two decades ago, the field of study called 'science, technology and society' began systematically to assess the ethical, legal and social implications of science and technology. Sociologists, ethicists and historians during the 1970s addressed questions of privacy, distribution and discrimination associated with such technologies as computer databanks and medical diagnostics, research practices such as human experimentation and organ transplantation, and scientific claims such as the relationship between crime and heredity or race and IQ. Most of their studies involved technologies already developed and issues already in dispute. By contrast, the founders of the Human Genome Project decided at the outset to invest at least three per cent of its annual budget in a programme called Ethical, Legal and Social Implications (ELSI).

*Gene Mapping* includes the papers from a January 1991 workshop convened to outline for ELSI the broad social policy issues that are likely to emerge from genome research. In an introductory essay, James Watson, former head of the genome project, and Eric Juengst, director of ELSI, tell us that genomic research will "provide insights into the treatment and prevention of both inherited disorders and diseases. . . [and] shed light on our evolution as a species and our development as individuals. And it

will serve as a starting point for researchers exploring biological elements in behaviour." The implications of such research and the potential misuse of information about genetic predisposition clearly call for responsible social policy. ELSI's important and worthwhile goal is to develop the basis of such policy as the science evolves.

The editors of this book compare gene mappers to Columbus, setting out to discover a new terrain, and present their book as an effort to assess the "voyage" of the Human Genome Project so as to "guide policy makers in choosing the directions that should be travelled in our genetic future". Accordingly, the chapters, written by philosophers, lawyers, historians and clinicians, examine the vast array of problems raised by this much discussed scientific endeavour.

Several essays describe the implications of genetic screening for clinical practice, noting the problems of privacy and control of information about genetic predispositions when insurers and employers as well as family members have a stake. Others call attention to the dilemmas of genetic screening for diseases that can be diagnosed long before there is any hope of cure. One lawyer writes on reproductive freedom, another on patent rights. One philosopher writes about the implications of reductionism and determinism, another on the implications of gene therapy. Many of the chapters touch on the threat of a new eugenics, although none suggests the form that this might take in the contemporary political context.

The book concludes with recom-

mendations, listing four questions that should shape ELSI's future research priorities. When and how should genetic tests be introduced into medical practice? How can confidentiality of genetic information be preserved? How can discrimination be prevented? How might the genome project affect our concepts of disease, normality and humanness?

The essays are excellent, the authors stellar — all scholars who have been writing on these issues for years. But by now, this volume is but one of many books, articles and edited collections dealing with the same broad range of issues. Indeed, they raise the very same questions that preoccupied those concerned about earlier technologies: the privacy of information in computer databanks; the personal dilemmas following from advances in clinical diagnostics that have always preceded therapeutic capabilities; the discriminatory use of tests; the age-old abuse of biology — "natural" categories — to explain and justify existing social relationships; and, of course, the notorious misuse of genetic ideas for eugenic goals.

It seems pertinent to ask whether 20 years of speculation about these problems has influenced the way we deal with advances in genetics. Has concern about the social implications of technology affected public understanding of issues such as privacy, created more sophisticated discussion in the media or resulted in more sensitive policy approaches? The record is not very promising. A recent survey exploring public attitudes towards genetic privacy found most people to be not very sensitive to the issues of confidentiality, and even willing to share genetic information (probably others', not theirs) with insurers and employers. Despite the increased use of genetic screening and awareness of the importance of genetic information, there remains a dramatic shortage of trained genetic counsellors. The media persist in oversimplified interpretations, describing the most complex human behaviour as 'programmed by the genes', treating biology as destiny and justifying gender stereotypes in terms of 'natural' traits. The appeal of reductionist explanations and technological fixes persists in the continued tendency to provide medical explanations of social problems. Some scientists themselves perpetuate misconceptions, insisting on the power of nature over nurture and promoting genetics as a cure for social pathologies as well as inherited disease.

Thomas Murray, one of the authors in *Gene Mapping*, makes a salient observation about the social policy discussions of the Human Genome Project: "Rarely, if ever, has something gone so swiftly from prophetic warning to cliché." The essays in this and other books have clearly been

helpful in laying out important issues. But if ELSI projects are to enhance public understanding and influence social policy, this is no longer enough. We need more concrete data addressing the institutional, political and cultural context in which genetic information will be used and abused. Historical lessons, philosophical inquiries, scientific wisdom will have greater impact if we understand the contemporary contexts in which science is applied. By now,

empirical investigations should be testing theoretical speculations, and scholarly explorations should be reflected in public education. More recent ELSI projects are directed to these ends and should soon be reflected in scholarly writing on the subject. □

*Dorothy Nelkin is in the Department of Sociology, 269 Mercer Street, New York University, New York, New York 10003, USA.*

## Sex and the singular woman

Richard Davenport-Hines

**Marie Stopes and the Sexual Revolution.** By June Rose. *Faber and Faber*. 1992. Pp. 272. £14.99, \$22.95.

MARIE Stopes is best remembered as the brave and generous woman who in 1921 opened Britain's first birth-control clinic, staffed by women and providing free advice. She was also a doughty sexual publicist whose manual *Married Love* (1918) "came crashing into English society like a bombshell", as its author described, with an "explosively contagious" denunciation of the habit "of using the woman as the passive instrument of man's need".

Born in 1880, she was a brilliant oddity: her father was an ardent amateur palaeontologist who stimulated her interest in science; her mother drove her into a formidable if unamiable precocity. By the age of 17 she had read all of Darwin, Kant and Swedenborg. She took science degrees at London and Munich universities and in 1905 became the youngest ever Doctor of Science. During the following year, despite being penniless, she took herself on an intrepid botanizing tour of Japan. She applied to join Scott on his Antarctic expedition, and was outraged to be rejected on the grounds of her sex. At considerable emotional cost, which she hid from herself, she attended meetings of scientific societies which were misogynistic and even contemptuous.

The course of her life was changed by the failure of her first marriage. Full of sexual yearnings but confused by romantic ideals, she married in 1911, but obtained an annulment in 1916 on grounds of nonconsummation (although June Rose suggests that Stopes's complaints of her first husband's impotence were untrue). Camping alone on a windswept beach at this time, she compiled a "Tabulation of Symptoms of Sexual Excitement in Solitude". It was in this state of sexual frustration and resentment, claiming still to be an unhappy virgin, that she wrote the book that made her

name, *Married Love*.

For hundreds of years there had been sexual manuals written for men. This was the first such book written by a woman for other women describing foreplay, sexual techniques and orgasm. "In it she dared to stake a claim for female sexuality", as Rose writes. "Her views



**The lovers' guide. Stopes in fact emerges as a magnificent monster.**

challenged centuries of prejudice and superstition and the accretions of religious teaching which saw women's bodies and women's attractions as desirable but also dirty." She wrote with an intensity of feeling to which other women responded, although the quaintness of some of her phrases can seem absurd today.

Thereafter she became increasingly sex-obsessed. She was a prolific writer and vivid controversialist whose campaigns involved her in lawsuits, ostracism and public abuse. She never shirked what she saw as her duty, and never spared herself. Her private life was turbulent. She married again in 1918, and was inconsolable when her first child was

stillborn. In 1924 a second son was born. At home (as in many of her public dealings) Stopes was tactless, humourless, cruel and headstrong, a crashing snob and driving bully who made an arid waste of the lives of several of the people nearest to her. She was closer to her vicious chow-chow Wuffles than to her immediate family.

June Rose writes in a tone that mixes admiration with irony. There is much to admire in the life of this energetic and daunting woman. The letters to Stopes from ill, exhausted women of the slums asking for contraceptive advice, or thanking her for her work, are exceedingly moving documents. But like other early popularizers of contraceptive education, she was interested in eugenics and developed a propagandist zeal for racial purity that seems ugly to later generations. Her lack of humour and her tunnel vision understandably antagonized medical opinion and should have served as a warning to sex educators ever since.

Stopes was sensitive to her status as a scientist, yet her contacts with the scientific community are slightly neglected. She suffered (perhaps more than Rose acknowledges) from the isolation and hostility that she met as a woman scientist in the early twentieth century. From the moment of her appointment as a lecturer in botany at the University of Manchester in 1904, the authorities were "watching her ... carefully", as Rose describes, in some cases hoping for her failure. She became even more exposed when her interests shifted from botany to sex education. Like her contemporary eugenicist Sybil Nevill-Rolfe, a sex educator and campaigner against venereal diseases, her tactics and manners varied from the irritating to the laughable. The unpleasantness of their position as women sex educators at a time of high male anxieties excited all their more unpleasant traits and more extreme tendencies.

Stopes became increasingly crankish and self-infatuated. There is much in this book that makes her seem supremely ridiculous: she was prone to gestures such as sending Hitler a copy of her poems, *Love Songs for Young Lovers*, or enthusiasms such as making contact with aliens from outer space (she claimed to have seen a flying saucer). Nevertheless she emerges as a magnificent monster, whose achievements were exceptional and enduring. She was a great fool and a great heroine. This is a readable biography of a scientific popularizer who changed the lives of millions and gave good service to her fellow women. □

*Richard Davenport-Hines is at 51 Elsham Road, London W14 8HD, UK.*



# Memoirs of a septuagenarian

Roy Porter

**The Diary of William Harvey: The Imaginary Journal of the Physician Who Revolutionized Medicine.** By Jean Hamburger. Rutgers University Press: 1992. Pp. 255. \$35 (hbk), \$14.95 (pbk).

WHO knows how many readers have read *The Diary of a Surgeon*, published in 1937, believing it to be the journal of the young medic, William Knyveton, and a reliable account of the horrors of eighteenth-century practice, without the slightest suspicion it was a hoax played on them by that wily medical writer, Ernest Gray? Any such temptation to tease has been quite wisely resisted by the distinguished French medical researcher Jean Hamburger in his overtly fictional reconstruction of the mental musings of Britain's most eminent physician.

Even so, this touchingly written and beautifully translated volume is still not what one might expect. The natural subject-matter for a fictional diary would have been a day-by-day account leading from Harvey's student days up to the thrilling events of the 1620s when he convinced himself that the Ancients had quite misunderstood the function of the heart and had fabricated a cardiac anatomy that was in crucial aspects utterly fanciful. Bravely challenging Classical dogma, Harvey had then undertaken the monumental series of experiments that demonstrated the circulation of the blood — even if, to Harvey's intense irritation, notable physicians such as Riolan in France insisted on upholding the *ex cathedra* judgements of Galen against the experience of the eyes. A blow-by-blow account of the great breakthrough would have proved an exhilarating read, offering Hamburger a golden opportunity to rethink the dialectic of hypothesis and experiment leading to *De Motu Cordis* (1628), and so perhaps to gloss modern conceptions of creativity and what Thomas Kuhn calls the "structure of scientific revolutions".

This is not what Hamburger has attempted. Rather, opening in 1647, he has given us Harvey in his seventies, gout-ridden and sometimes crotchety; and he has imagined not so much a diary but rather the memoirs of a man of full years, still obliged to defend his discovery, and not just against the Antidiluvians but also against those new men, such as his crony Thomas Hobbes and also René Descartes, who were putting it to use in mechanical philosophies that aroused deep suspicion. Harvey's child was breaking away from its father.

It is a subtle choice. It gives Hamburger the opportunity to depict Harvey as a mature thinker, complete with the philosophy of organic life more evident in later embryological writings such as *De Generatione* than in his studies of the heart. Hamburger's is a respectful Harvey, who can cherish the wisdom of Aristotelian teleology, while repudiating the medical old guard's Galen-worship. It is a Harvey who has honed natural philosophical maxims of his own: "doubt is the most important quality a scientist can possess".

By choosing the septuagenarian, Hamburger can explore Harvey-the-man no less than Harvey-the-researcher. For many years King Charles I's personal physician and growing politically more crusty with age, the Royalist Harvey lamented the Great Rebellion. The year 1647 found the country in limbo, the Civil War over, the situation of King, Army and Parliament unresolved. Through a clever device, the great shift in the three years of diary here 'recorded' is not some event in Hamburger's hero's life *per se*, but Harvey's response to the unforeseen, unthinkable but, with the aid of hindsight, inexorable *via crucis* of the king towards the scaffold in January 1649: an event that reduces the normally garrulous physician to pathetic near-silence: "2 February 1649. The King is dead. His death is murder and, even though he remained in the background, I hold Cromwell to be his murderer." It is a neat irony that Harvey, who had earlier dethroned the heart, reducing it from being the very monarch of life into a mere muscle, a mechanical pump, should nevertheless have deplored the parallel revolution within the body politic.

Hamburger has a flair for evoking the strange pageant of Baroque Europe, half-old, half-new, one foot still in the camp of magic and the occult, the other in the world of Galileo and the new philosophy: the Rosicrucian Robert Fludd, Gassendi, Mersenne, John Evelyn and John Milton (as politician not poet) all flit in and out of the pages. There is a memorable vignette of the courtier Sir Kenelm Digby touting mountebank-like his unique "powder of sympathy" that supposedly healed wounds by sympathetic magic, while sustaining his beautiful wife Venetia only on capons fed on viper flesh, hoping to preserve her youth. Hamburger's reconstruction of Harvey's life and mind combines accuracy with an allowable imaginative licence. Here and there, it is a



**In 1616, the year Harvey first lectured on the circulation of the blood, Nicholas Culpeper, popular figurehead of 'alternative' medicine and Harvey's polar opposite, was born. With his unauthorized translation of the *London Dispensatory*, Culpeper became the enemy of the physicians, and in 1652, he wrote his famous herbal *The English Physician*. The myths surrounding his life are now explored by Olav Thulesius in *Nicholas Culpeper: Physician and Astrologer*. Macmillan, £40.**

bit too knowing ("... this epoch is creative. Everything will progress very quickly from now on"). Yet in general, Hamburger admirably catches the solemn, melancholy mood of a greybeard oppressed by the vanity of all-talk no-action intellectuals such as the despised courtier Francis Bacon, who, quipped Harvey, did science like a Lord Chancellor. Harvey was committed to belief in scientific progress — there lay the only escape from the bloodshed and butchery that had attended all his days. Yet even towering medical advance led the old man's mind back to an inescapable anatomy of melancholy: "my eyes will never forget Fabricius' Anatomy Theatre; at the bottom of the five circular tiers, on the central dissecting table, a corpse, the anatomical model, showed us all what becomes of a man at the end of his earthly history".

This is a *jeu d'esprit*. It is also a splendid attempt to probe beneath the skin of a figure, the sheer range of whose biomedical achievements is all too rarely appreciated. For Hamburger's Harvey, pursuit of truth was the pole star, the equivalent to the sexual drive in others. Avoiding the pitfall of twee archaic language, *The Diary of William Harvey* is also, as the extracts show, an enchanting read. Hamburger's death earlier this year highlights the poignancy of his moving book. □

Roy Porter is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BN, UK.

# Improving our minds

Mary Warnock

**Education Without Impact: How Our Universities Fail the Young.** By George H. Douglas. *Birch Lane*: 1992. Pp. 223. \$19.95.

THERE are two main threads in George H. Douglas's highly critical analysis of universities in the United States. The first is that these universities are dominated by science, the second that, in his words, "Americans have gravitated toward the idea that the university is like a giant department store . . . a place where people come to shop for things."

The domination of the natural sciences has led, in Douglas's view, to a corruption of the humanities. He claims that because the sciences have inevitably become increasingly specialized, and because there has thus grown up a series of specialist languages embodying concepts barely intelligible to the general public, professionals in the humanities have felt that, to retain their respectability, they must follow the same path. The studies of literature, philosophy, history and languages have all become pseudo-scientific, no more accessible to the lay person than science itself.

Especially in the case of literature, Douglas, himself a professor of English, argues that theories of criticism have been developed over the past 40 years or so that have become more important within university departments than the literature they were originally supposed to illuminate. (No one familiar with the English faculty in the University of Cambridge would be much inclined to dispute this.) And just as scientists are concerned above all with research, and their interest, if any, in teaching is directed towards producing new researchers, so in the humanities research has taken precedence over teaching, the only worthwhile teaching being of those pupils who may themselves take up research. The idea of passing on a civilized and civilizing tradition has disappeared. The Germanic concept of the learned doctorate, the scholarly DPhil, is blamed for this degeneration as much as the rise of science itself. But once the DPhil is firmly in place, it is hard to eliminate, because it becomes a necessary step to advancement. Such corruption of the humanities has been a creeping phenomenon, and has, according to Douglas, been visible all through this century.

The development of the university as a supermarket is of much later growth, although the two phenomena are not unconnected. According to Douglas, by the 1960s, undergraduates had become not only numerous, but also too patently neglected by their professors, overly

accustomed to being fobbed off with huge impersonal lectures and to having their examinations graded by graduate students scarcely older than themselves. So it was inevitable that they should decide to take matters into their own hands and demand to change the university curriculum to make it easier and more user-friendly. They also began to demand, as customers will, better value for money; and this meant higher grades. So while the test grades of school-leavers became lower, the grades they were given in their university examinations grew ever higher. They had come to think of undergraduate education as simply a stage to be got through before going on to the graduate or professional stage; and to ease their passage they needed good grades.

Everyone reading this threnody will find a good deal in it that is recognizable; it is a threnody, moreover, that has been sung before. It is inevitable that readers should compare what has happened in the United States with what has happened, or is about to happen, in other countries. The comparison with UK universities in particular is absolutely irresistible. For Douglas is an Anglophile, at least as far as university education goes. Moreover, he is nothing

if not romantic. He draws a picture of the community of scholars in the colleges of Oxford and Cambridge fit to wring tears from a stone. He is perhaps unaware that the last but one UK Secretary of State for Education, Kenneth Baker, advised five years ago that British universities should learn to model themselves on the American pattern. Since then the number of universities has vastly increased, and the overall number of undergraduate students is rising, but with no more resources for teaching.

Douglas accepts, perhaps too readily, that the universities of Oxford and Cambridge are elitist institutions, and that their style of undergraduate teaching depends on relatively small numbers; but he takes a tough line about this. Elitism is, in his eyes, a necessary good for universities. Nevertheless he is insistent that professional academics must not get out of touch with society at large, their contacts being their own pupils (whom they are to meet in the idyllic setting of the college) and their publications. They must be proud to write for popular, at least readable periodicals, and generally make use of the media to disseminate their ideas.

This is sensible advice, as far as it goes. But there are two major issues that today preoccupy everyone concerned with higher education, and on which Douglas is completely silent. How is the idea of a university to be reconciled with the idea of near-universal tertiary education? And how are universities to be funded? The absence of any discussion of these issues gives his book a curiously old-fashioned flavour. It seems to me



**Collected collections** — the Dutch taxidermist William Cornelius van Heurn (1887–1972) preserved a huge number of animals to illustrate the variation within species that drives evolution. This picture of some of his moles appears in *Finders, Keepers*, a survey of eight natural history collectors and their strange quarries, with marvellous colour photographs by Rosamond Wolff Purcell and accompanying text by Stephen Jay Gould. Hutchison/Norton, \$50, £19.99.



that the logic of his position is that only a handful of students will benefit from what he regards as true university education, whether in the sciences or the humanities. Only a few can be taught face-to-face by tutors who know them and are prepared to discuss their work with them in detail. Others will need technical expertise, or perhaps preliminary training, before going on to law school, business school or medical school, or into employment. This would entail a distinction, acknowledged or otherwise, between true universities, essentially suitable for only a minority, and training colleges, rather like the original distinction between universities and polytechnics (a distinction which is now defunct in the United Kingdom). It would also entail a distinction between academically indulgent institutions, with high and general educational aims, and functional institutions with precise practical goals, expected to give clear value for money. The likelihood is that public funding would be available only for the latter. But perhaps Douglas would not mind. Rightly or wrongly, the United Kingdom has just turned its back on this kind of dichotomy. For the time being, all universities are supposed to be equal.

There is, however, one further question, in my view the most fundamental of all, to which Douglas's answer is unclear. What is the ideal relation between teaching and research? He blames professors for taking their research too seriously, and palming off their teaching on their juniors. But he does not suggest any means to reverse this practice. In the United Kingdom it begins to look increasingly as if research, especially scientific research, will come to be based in a few selected universities, undergraduate teaching being the main responsibility of the rest. Graduate students will presumably find their way to the research universities, to be supervised by those actively engaged in research. Undergraduates will thus be taught by professional teachers, as they were at school. The teaching institutions will become, in effect, tertiary schools ('schools with ash-trays', as someone has called them). Professors will no longer incur blame for not caring about undergraduates, because they will have no undergraduates to care about in their research-orientated lives. Things may well come to this. When they do, I shall almost feel inclined to join Douglas in his nostalgia for the old days at Oxford and Cambridge, though I feel sure, as he apparently does not, that those days are over (and in any case, they were never perhaps quite so glorious as he assumes). □

Baroness Warnock is at Brick House, Axford, Wiltshire SN8 2EX, UK.

## Unreasonable ambition

Ernest Gellner

**Irrationality: The Enemy Within.** By Stuart Sutherland. Constable: 1992. Pp. 357. £14.95.

STUART Sutherland is a rationalist. So, as it happens, am I. So you might expect me to acclaim an author who expresses views, attitudes and values very much akin to my own, and who does it with verve, eloquence, coherence and in good, clear prose. In fact, however, this book leaves me feeling very uneasy. Why do I not salute an intellectual friend and ally? Why the reservations about this survey of modern follies and their causes?

Let us go straight to the heart of the matter. Sutherland is not merely a rationalist, he is also a psychologist. Moreover, although he is clearly a highly intelligent and able member of the species, the trouble is that he is also, in his basic attitude, a fairly typical one. Scratch an anthropologist and you'll find a philosopher, and sometimes even one with a sense of history; sociologists sometimes have the same merits; but psychologists, seldom. Scratch a psychologist, and you'll generally find a man eager to flee the humanities and act up to a certain conventional model of a scientist. I'm not sure this is true of all the other books of Sutherland's, notably *Breakdown*, which was highly personal, but it is true of this one.

It isn't that I in any way disapprove of Sutherland's aspiration to expose irrational behaviour: my worries spring rather from the fear that he has greatly underestimated the difficulties of the task. The underlying assumption of this book is that its subject's opposite, reason, is easy to identify. It is, as Spinoza said about truth, the touchstone both of itself and of its error. It is readily available, luminous and obvious, and so is unreason. I do not know whether Sutherland would plead guilty to this accusation when so spelt out, but its influence on his argument, at any rate in this book, is evident from the terrible clarity of the battle lines: he can spot unreason so very clearly; he knows his enemy, because he also knows, with disturbing and facile confidence, how to identify *our* position, rationality. Neither poses any problems for him, really.

His conception of rational thought and conduct owes a great deal to basic statistical ideas and to the economists' conception of sensible market behaviour. Rational man is careful in his handling of evidence, avoids well-known fallacies

and rigidity and sheep-like behaviour, and his lucidity in making inferences from the database at his disposal is matched by his clarity about his aims and the assessment of what will serve them best. Clearly he is an enlightened individualist who will not follow a multitude to commit folly.

An admirable ideal, and I share Sutherland's admiration. But the book contains no serious discussion of the social and historical preconditions of the emergence of this strange animal, or any sense of his untypicality, in wider historical context. The index does not contain the name of Max Weber or of any of those who have been puzzled by the emergence of rationality, not even that of James Frazer (whose views on this topic were not so far from Sutherland's, though he had far more sense of my feeling for the poetry of unreason). The disaggregation of the world into isolable elements that can then be subjected to statistically proper interpretations, the disaggregation of conduct into isolable and freely chosen aims permitting cost-benefit calculations — all this is the achievement of one type of society, it is not inherent in the human condition. It is not part of our birthright. Like believers in a Revelation who condemn to hellfire those who preceded it and had no access to it, Sutherland's ideal implicitly condemns much of mankind to servitude to unreason. Perhaps he would say that there is no point in deploring that which is beyond remedy, but the problem is not perceived.

But the situation isn't all that simple even among us, where reason is an important though not exclusive norm. Another obvious name missing from the index is that of Thomas Kuhn. Without accepting everything in the work of this philosopher, it is difficult to dispute his main claim about the nature of science: namely that, except in the course of major revolutions, science works and can only work through a kind of socially, culturally induced dogmatism, a pre-judgement of major and objectively contentious issues, the imposition of a shared paradigm. If this is true among scientists working in a milieu disconnected from their social identity and its pressures, and in the main imbued by an individualistic-rationalistic spirit, how much truer it must be of ordinary life and its strategies, which are Sutherland's concern! To put it another way: the data at our disposal do not uniquely dictate what we should think about our environment. Far from it. Theories, as the phrase goes, are under-determined by facts. There is simply far too much slack between evidence and conclusion: unreason, sheep-like thinking, rigidity, call it what you will, must take up this slack.

To say all this is not to ask Sutherland  
NATURE · VOL 360 · 26 NOVEMBER 1992

to write a different book from the one he set out to write. He has indeed taken our current notion of rationality for granted, without probing deep into its basis, and gives us an interesting survey of the violations of this norm. But he also aspires to go further: at the end of the book, he offers a list of the hindrances to rationality, which by implication seek out its roots. The list runs as follows: our evolutionary heritage, our neurological equipment, laziness, ignorance of statistics and self-serving behaviour. A curiously heterogeneous set of basic sins . . .

Apart from lacking historical sense in broad outline, the author is exceedingly slapdash in handling specific facts. He likes to invoke examples of military folly, but tends to get it wrong. On page 41, the Light Brigade in the Crimea is described as charging Turkish (*sic*) guns and soldiers. Lord Raglan may indeed have been "doltish" but he could tell enemies from allies. Sutherland seems to have little sense of geography: if the Turks had indeed been the enemy, how on earth could the expedition have ever passed through the Dardanelles and reached the Black Sea? On page 144, General Montgomery is similarly derided for failing to take Antwerp in 1944, thereby enabling the German 23rd Army to escape from northern Holland (*sic*) and help defend Arnhem. The Germans had no need to escape from northern Holland (they stayed there, undisturbed, until the end of the war), and even less need of Antwerp to reach Arnhem from there, as Antwerp is nowhere near the way from north Holland to Arnhem. (Sutherland means northern Belgium.) Those soldiers may be fools but they do occasionally consult the map. The ends of each chapter of the book contains pithy advice on how to avoid folly. I commend one addition: when telling stories about the folly of others, check your own account for howlers.

All the same, this is a lively and readable book about modern follies, even if it fails to tackle the harder questions about the role of reason and unreason in human society and the human psyche. □

Ernest Gellner is at King's College, Cambridge CB2 1ST, UK.

### 1993 review supplements

Nature's review supplements next year are Spring Books (15 April), Autumn Books (25 November) and New Journals (7 October). The latter will cover journals launched during or after June 1991 with at least four separate numbers issued by the end of April 1993.

## La neige d'antan

Walter Gratzer

**C. P. Snow and the Struggle of Modernity.** By John de la Mothe. *University of Texas Press*: 1992. Pp. 288. \$35.

WHEN the first laboratories sprang up in Victorian Oxford, the wife of the Warden of All Souls observed with contempt: "The Warden could get up science in a fortnight if he wanted to". Seventy or so years later, so C. P. Snow tells us, the mathematician G. H. Hardy reflected: "It's rather odd, but when we hear about



Snow — poor spokesman for science.

'intellectuals' nowadays, it doesn't include people like me and J. J. Thomson and Rutherford". Snow's career as writer and magus was rooted in his indignation at such affronts to the first of his chosen callings. He vociferated it loudly and often: "Not to have read *War and Peace* and *La Cousine Bette* and *La Chartreuse de Parme* [in the original no doubt] is not to be educated, but so is not to have a glimmer of the Second Law of Thermodynamics." It was probably the intolerable reek of humbug that so got up the noses of F. R. Leavis and other literati of the day and whipped them into such passions of wrath.

I happened to be at Harvard when Snow, just then at the peak of his unaccountable fame, came to deliver the Godkin lectures (later published as *Science and Government*). His audience, mainly of students from Harvard and MIT, filled a theatre the size of a baseball stadium, and he basked in their adulation. Many had already sent off for his collected *oeuvre*, bound in imitation morocco with genuine 18-carat gold-leaf lettering. Here, plainly, was the stuff of future PhD theses.

But that was 30 years ago, and I had not supposed that anyone bothered too much with Snow any more: yet now John de la Mothe has dished up for us his densely written and minutely researched treatise, bearing on the dust flap beneath the portentous title the familiar image of Lord Snow of The Two Cul-

tures, turning on the world the morose gaze of a costive bulldog. "The discourse of modernity is comprised of a cacophony of voices, the interpretation of which can only be described as a struggle." Thus de la Mothe, getting his book off to an unenticing start. More, "it should not be surprising that our aspiration to more precisely delineate the parameters of this struggle has become the paradigmatic idea of our age." Well include me out when it comes to delineating parameters. Fighting down an urgent desire to bolt for cover, I ploughed on through the waterlogged terrain, feeling dry ground underfoot only when de la Mothe began to concern himself with Snow's life and especially his erratic scientific career.

De la Mothe identifies Snow with the thrust towards social progress and intellectual liberation — modernity, as he calls it — and while by no means uncritical of his subject, he accepts him substantially at Snow's own valuation of himself as the Messiah of the scientific age, both in his didactic utterances (the Rede and the Godkin lectures) and in his novels. These of course are peopled almost exclusively by scientists and academics and political mandarins, who stalk the Corridors of Power, gravid with the authority of their creator's years as scientist, civil servant and (briefly) politician. *Science and Government* purports to be a documentary study of decision-making processes in government, but as Snow remarks of his main protagonists, the deplorable Frederick Lindemann (Lord Cherwell) and Henry Tizard, they made the novelist's fingers itch; and Snow scratched where it itched (which, it must be conceded, makes the story an entertaining read, just as the novels often are).

Snow claims for scientists unique access to a wordly wisdom that few would arrogate to themselves (let alone their colleagues). They have, he asserted, "the future in their bones", a faculty denied by implication to members of other professions. To my mind, he was a poor spokesman for our trade. As a novelist he was pretty good for a scientist, as a scientist better at least than most novelists and as a politician merely, by all accounts, a failure. Lewis Eliot, the narrator of the *roman fleuve*, *Strangers and Brothers*, though a lawyer, is Snow himself — supercompetent, knowing and superior, the reflection in Narcissus's mirror. What undoubtedly lends spice to the novels is that we can identify (for Snow made no attempt to conceal them) the models for the characters, and to a certain extent the events also touch on reality. The trouble is that for all the solemn air of authority in which his characters envelop themselves, their aspirations are for the most part essen-



tially trivial (as in *The Masters*), just as Snow was himself an inveterate pot-hunter and seeker after life's baubles.

The most interesting of the novels remains *The Search*, for it describes, often with compelling immediacy, the people and episodes that directed the course of Snow's academic life. The hero, Arthur Miles — again Snow — is thwarted in his purpose to become director of a new biophysics institute by the untimely discovery of a fearful gaffe in the work that has made his name. Crushed by disappointment and humiliation, which his professional rivals do nothing to assuage, Miles rejects the advice of a friend to apply himself patiently to his *métier* and rebuild his reputation, and resolves instead to forsake science and seek fame as a Man of Affairs. What a piffling reason, J. B. S. Haldane commented, for giving up the privilege of a life in research — the mere frustration of a personal ambition! But Snow was first and last an *arriviste*, whose interest was not in the journey, only in the goal. His own research, of which de la Mothe gives an absorbing account, was an almost unbroken succession of boners, the consequence most often of intellectual delinquency or wanton carelessness. The climax came with the publication by Snow and Philip Bowden (Francis Getliffe in the *Strangers and Brothers* series) on the spectroscopic identification and supposed photochemical generation of vitamin A. Their claims were brutally anatomized by Ian Heilbron and R. A. Morton. In particular, and not for the first time, Snow had made a fool of himself by his ignorance of the literature; he was probably already too busy with his other ambitions to bother with reading journals.

The metamorphosis of Percy Snow (as he was known to his fellow-students, who found him such a pest, in the chemistry laboratories at Leicester) into Sir Charles and later Baron Snow, his rise to the status of sage to that unsettled decade, the sixties, and the curious growth of his literary reputation make an interesting moral tale that repays study. Snow went far on his catch-phrase, *The Two Cultures*, but the idea was not new: Waddington, Koestler, Bernal, Polanyi and many others had articulated the same concerns, as in an earlier generation had H. G. Wells, a better writer by far than Snow and an altogether deeper intellect. That grandee of the other culture, Max Beerbohm, spoke of Wells's prose as "cold rice-pudding, spilled on the pavements of Gower Street". Why, then alas for Snow. □

Walter Gratzer is in the MRC Muscle and Cell Motility Unit, King's College London, 26–29 Drury Lane, London WC2B 5RL, UK.

## Imbalances of power

Lawrence Freedman

**Closing Pandora's Box: Arms Races, Arms Control and the History of the Cold War.** By Patrick Glynn. *Basic Books*: 1992. Pp. 445. \$30.

DURING the 1980s, US arms-control policy came under the increasing influence of a collection of neo-conservatives who were highly dubious about the whole exercise. Many of them, including Richard Perle in the Pentagon, had been

homicide squad. They were therefore somewhat surprised that the decade concluded with a spate of dramatic breakthroughs in arms control. The main reason for this was the collapse of Soviet inner strength and self-confidence, which led to an almost craven acceptance of whatever the Americans happened to be proposing at the time. This raised questions about the relationship between attempts to regulate armaments and the overall balance of power.

The old liberal hope was that by containing the arms race, not only would the danger of war by accident or miscalculation be reduced but so too would a major source of aggravation in political relations. By the 1980s one did not have



**End of the road — the collapse of Soviet power can be traced to the cumulative inefficiencies of state socialism, as much as to any external pressures from the West.**

previously associated with Senator Henry Jackson. They argued that their predecessors had paid far too much attention to 'negotiability' to the neglect of substance, so that deals had been applauded even when they were to the detriment of the West's security. By acting as if the details of the military balance did not matter, they had allowed the balance to tilt dangerously to the Soviet's advantage. Reflecting President Reagan's own views, they did not put a high premium on 'negotiability' when they did get around to negotiating.

Liberals thought that the prominent position of this group in arms-control posts in the Reagan administration was akin to putting Crippen in charge of the

to be a conservative to appreciate the problems with this perspective. In practice, when relations were tense, arms-control efforts tended to make matters worse, through, for example, arguments over compliance; when relations were improving, arms control at best reinforced the benign trend. Against the underlying liberal assumption that there was little other than mutual fear and suspicion driving the great powers to war was the unfortunate reality that the East-West confrontation was not simply an unfortunate misunderstanding but reflected real differences in ideology and interests.

The neo-conservatives took up this 'realist' position, but then turned it into

NATURE · VOL 360 · 26 NOVEMBER 1992

a dogma of their own. Their view was that when status quo powers such as the United Kingdom and the United States are faced with a radical power determined to upset the established order, they must pay attention to the military balance. If the liberal democracies allow their guard to fall because of fond hopes that restraint will be reciprocated and a spirit of goodwill created, they will simply find themselves at a chronic disadvantage should the crunch arise.

Patrick Glynn worked in the Arms Control and Disarmament Agency from 1986 to 1989 and is now at the American Enterprise Institute, a conservative think-tank. In this book he provides the most developed version of the realist thesis. It is much more than a polemic or a memoir (of which there have been a number recently). It is a well-researched analysis of the history of attempts to use controls on weapons to ease political tensions.

Glynn starts with Sarajevo — the 1914 rather than the 1992 version. The start of the First World War is often cited as the classic example of how catastrophic wars can be the product of arms races (in this case the Anglo-German naval competition) and mobilization plans that allow little time for effective diplomatic management of crises. Drawing on the weight of contemporary historical research, he demonstrates that the basic problem was the German determination to overturn the status quo and that deterrence failed because expressions of liberal sentiment gave German hawks hope that Britain in particular would stay on the sidelines. He then follows this through by looking at the more familiar story of the paths full of good intentions that led to the Second World War, and then the shift of gear as the Western democracies came to terms with what they judged to be the long-term intentions of their erstwhile ally, the Soviet Union.

In the process of all this, Glynn marshals his evidence effectively, makes several novel and interesting points, and writes clearly. The difficulties with his approach lie more in the second part of the book where he applies this analysis to the Cold War. He assumes that power politics have continued to be in play throughout this period. Thus he takes the variations in the military balance to be of considerable importance, to the point where the collapse of the Soviet Union is traced to Gorbachev's stark realization that he could not cope with the challenge posed by Reagan's military build-up, including the Star Wars programme.

Here the narrative becomes too synoptic and the context too blurred. By stressing the continuities with the past, Glynn ends up missing the discon-

tinuities. The power politics of the first half of this century was followed by a period in which the political order congealed, at least in Europe. Without any real threats of war, the nuances of the military balance became far less important than was claimed by both the hawks and the doves.

The Soviet Union ceased to be a radical power but instead developed its own stake in the status quo and feared US radicalism. But the key instrument of this radicalism was not the arms race, but the strength of the economic system it represented and the political values it espoused.

Ironically, given the conservatives' distaste for *détente*, it was the stress on human rights inserted into the 1975 "Final Act" of the Conference on Security and Cooperation in Europe that eventually undermined the Soviet system, as much as any other external pressure

imposed by the West. Glynn does not mention this conference, perhaps because he finds it difficult to accept that anything good could have come out of the Kissinger era, though he acknowledges the importance of the West's demands for human rights. When Gorbachev began to face the need for reform, he acted to gain Western support by offering concessions on the key areas of Western concern — arms control, regional conflicts (such as Afghanistan) and human rights. In each case it suited his own purposes. But what finished him was an area in which no demands had ever been made — the cumulative inefficiencies of state socialism, which had reached the point where they were beyond reform. □

*Lawrence Freedman is in the Department of War Studies, King's College London, Strand, London WC2R 2LS, UK.*

## Promethean prescription

*Timothy O'Riordan*

**Green Delusions: An Environmentalist Critique of Radical Environmentalism.** By Martin W. Lewis. *Duke University Press: 1992. Pp. 288. \$24.95, £21.*

FOR Martin Lewis, radical environmentalism rests on four essential postulates, namely that primitive or primal people exemplify how to live in harmony with the natural world and with each other; that thorough-going decentralization based on autarchic self-reliance is necessary for ecological and social health; that technological advance, if not scientific progress itself, is inherently harmful and dehumanizing; and that the capitalist market system is inescapably destructive and wasteful.

In its place he advances the cause of promethean environmentalism, arguing that enlightened sophisticated people can learn to come to terms with nature and their own existence; that cooperation based on trust and scientific understanding should underpin future government; that technological development can be made environmentally benign if the right signals are put in place; and that a socially guided, more strategically caring capitalism is vital for the transition to sustainability. A healthy society, he concludes, is one characterized by simultaneous advances in general prosperity, social equity and environmental stability.

The text itself is a well referenced analysis of the radical view and a less convincingly argued case for prometheanism. But Lewis falls into the trap of setting up stereotypes to represent

what are really fragmentary minority positions. Few intelligent analysts of environmental values espouse what Lewis characterizes as radical environmentalism. True, there are some biggish names such as Murray Bookchin, Rudi Bahro and David Pepper who write about deep greenism and the sociology of the commune. But the heartland of modern environmental theory is closer to Lewis's position than he is prepared to admit. Indeed, his analysis smacks of a flexible and adaptable environmentalism that most social democrats, or old-style liberals, already proclaim as their territory.

In many ways, therefore, Lewis is tilting at windmills when the real target lies uncomfortably in his own stance, itself dangerously off balance. No one seriously believes that any of the four postulates is either 'true' or a blueprint for the future. Even the most woolly of the Green Party activists see a role for benign technology, for some sort of social contract with commerce, for carrying degrees of national and international intervention in the otherwise corrupted capitalism of the modern environmental age, and for ensuring that indigenous peoples have dignity as well as a degree of political independence.

Nevertheless, Lewis's argument is a goldmine of thoughtful assessments of a wide range of social reasoning in the modern environmentalist debate, and he incorporates text that will probably be new to even the most well read of readers. Of special value are his insights into the relationship between capitalism, the state and the longer-term public purpose. Here is one arena where mod-



ern environmentalism may well be creating genuine social and political change.

After the United Nations Conference on Environment and Development in Rio de Janeiro earlier this year, it is evident that neither democracy as we currently know it, nor international diplomacy as presently practised, is ready for the constellation of adjustments that will be needed to form the simple but ungraspable trio of ecological health, social justice and guided enterprise on a degrading world stage. The problem for Lewis is that he devotes most of his effort to knocking down the straw people he has comfortably but misleadingly created for himself, and said too little about how to shift to a sustainable future based on the structures that plague us today.

In the area Lewis calls guided capitalism, there is a slim chance of progress. The major multinational companies have the capability and power to set an agenda for more-sustainable capitalism on a global scale. It would require a heroic shift in pricing, self-regulation, openness and management, and it would mean greater cooperation in terms of promoting broad principles and codes of practice than anything yet envisaged. But it is in this arena of international guidance and multinational support that the beginnings of a transformation could take place. Success would depend, however, on interventionist and radically equalizing parties at the head of government, and a citizenry that treated sharing seriously. There is absolutely no evidence that such ingredients are in place. The best one can hope for is a public debate around the preparation of national plans towards sustainable development that should emerge out of the Rio process. How far it will do so remains to be seen, and for now one must be very pessimistic. But as the world grinds into greater calamity and as each fresh generation takes the triple alliance of ecological health, social justice and guided enterprise more seriously, out of many ashes some form of modified phoenix, not Prometheus, could still arise. But it would be a leaner and much more devastated humanity that would have to adopt the mantle. □

Timothy O'Riordan is at the School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

## Environmental dictionaries

- *Dictionary of Environmental Science and Technology* (revised edition) by A. Porteous. Wiley, £9.95 (pbk).
- *Dictionary of Global Climate Change* compiled by W. J. Maunder under the auspices of The Stockholm Environment Institute. UCL Press, £19.95.
- *The Earth Report 3: An A-Z Guide to Environmental Issues* edited by E. Goldsmith and N. Hildyard. Mitchell Beazley, £9.99 (pbk).

# Humble homily

N. A. Mitchison

**Pietà.** By George Klein. MIT Press: 1992. Pp. 297. \$24.95, £22.50.

GEORGE Klein will be well known to many biologists for his lifetime of research on cancer viruses as the founding director of the Department of Tumour Biology at the Karolinska Institute in Stockholm. His insight, pertinacity and imagination led him to create out of this specialized theme a structure of ideas that has deeply influenced contemporary immunology and developmental biology. The same qualities illuminate this, his second collection of memoirs and reflections (another is on the way).

Excluded from the executive arms of the state, the Jews of middle Europe gravitated towards intellectual pursuits. Their way of life was profoundly German in its orientation, and its destruction in the Holocaust is one of the ironies of history. Klein's upbringing in pre-war Budapest left him familiar with the main European languages, and with a love for the poetry of Rilke, Goethe, Baudelaire and (more curiously) Poe combined with respect for the ideas of Schopenhauer, Freud and the young Heidegger. He recounts the walks through the old city, the picnics on the island at its centre, the concerts and the fervid cafe discussions.

In 1944, at the age of 19, Klein escaped from the death train taking him from the city; many of his family perished. This book does not belong to the literature of the Holocaust, but its view of human nature is deeply influenced by that experience. It lets him sympathize with the poet Attila József, who dissociated himself from the bourgeois life of inter-war Hungary and who was condemned by others as schizophrenic. It underlines for him the desperate difficulty in the AIDS epidemic of changing human behaviour, especially in anything so basic as sexual practice. It prevents any surprise that the Central Intelligence Agency in the McCarthy era carried out experiments with LSD on volunteers.

Klein's main theme concerns the different ways in which individuals confront their own suffering and alienation. In some this generates creativity, a connection that is explored in some depth here, particularly in the instances of József and Nietzsche. The argument is of course not unfamiliar, and no doubt it needs the leisurely pace and breadth of an Edmund Wilson to do it proper justice; but the author's personal experience gives this account a special poignancy. The final chapter, from which this collection takes its title (loosely translated as

'compassion'), rounds up the argument and places it in the context of German philosophical writing.

Appreciating creativity in literature requires a knowledge of the language, a forlorn hope for most of us when that is Hungarian. Klein does his best, for instance in discussing József's provocative choice of a verb translatable as 'applaud' to describe what jasmine petals do. He succeeds within the limited space that he permits himself, and I hope will expand this line of thought in his next collection.

Klein tells a good anecdote. We learn in detail about an afternoon's conversation between François Jacob, Bruno Müller-Hill, himself and the now Canadian neuropharmacologist Rudolf Vrba. Vrba had participated in the first escape from Auschwitz, and had jointly written a report on conditions there that did much to save lives among the Jews of Budapest. These men's argument about the duty of the scientist, and the question of whether science itself has an internal ethic, will interest any reader of *Nature*. Klein is not at all egocentric, and in fact rather amusing about himself: totally confounding the officialdom of the airline Swissair, ferreting out the personalities that interest him, and defeating the efforts of his fellow citizens in Stockholm to engage him in chit-chat.

This collection also includes a little science. The brief histories of T-cell biology and of the AIDS epidemic will interest nonspecialists, but Klein's very lack of egocentricity leaves one unsatisfied. I would have liked more about how he felt at the time and what impression the rest of the community made on him.

One of the things that Klein is best known for is the tape recorder that invariably accompanies him on his aeroplane flights. Well, anyone who wishes to emulate him in that respect should leave this volume behind: once started, it is impossible to put down. More generally, no science library that aspires to more than the narrowest interpretation of its function can afford not to purchase this volume to place alongside its Jacob, its Müller-Hill and its Medawars. And a final note to the professionals: the European Community plans next year to make awards in the field of medical ethics, in the hope no doubt of catching up with Golub's Pacific Center for Ethics and Applied Biology. This volume, and its predecessor *The Atheist and the Holy City* (MIT Press, 1990), in my opinion make ideal background reading for those thinking of submitting applications (the closing date for which is January 1993). □

N. A. Mitchison is in the Deutsches Rheuma Forschungszentrum, Robert Koch Institute, Nordufer 20, 0-1000 Berlin 39, Germany.

# The emperor's old clothes

Timothy Ferris

**From Eros to Gaia.** By Freeman Dyson. *Pantheon*: 1992. Pp. 371. \$25.

MANY eminent scientists, having attained acclaim through their research in physics, mathematics or some other specialized discipline, become eager to write on matters far beyond their fields of expertise. The resulting books and essays, obedient though they may be to Schopenhauer's encouraging dictum that "the first forty years of life give us the text [while] the next thirty supply the commentary", are not uniformly praiseworthy. Indeed, to be blunt, the popular science literature is littered with windy, ill-thought-out tomes written by scientific celebrities who honour publishing contracts by hastily committing their loosely anchored opinions to the printed page. Particularly notable are 'follow-up' books — collections of speeches, old essays and juvenilia siphoned from the files of scientists whose earlier *haute vulgarisations* did well enough for their publishers to urge them to get another volume out pronto.

Freeman Dyson is an eminent physicist whose popular books (*Infinite in All Directions*, *Disturbing the Universe*, *Weapons and Hope*) are widely read on both sides of the Atlantic. *From Eros to Gaia*, his latest offering, scrapes so deeply into his filing cabinets that the merest perusal of its table of contents is apt to awaken a sense of alarm among students of the genre. It begins with the unfinished manuscript of a science fiction tale, "Sir Phillip Robert's Erolunar Collision", written by Dyson in 1933, when he was nine years old. Most of the rest of its 35 chapters are based on speeches Dyson gave and book reviews and essays he contributed to *The New Yorker*, *The Mathematical Intelligencer*, *Scientific American*, *American Journal of Physics* and other publications.

Yet to read this book is to dispel one's trepidations. Dyson turns out to be a happy exception to the baleful rule.

Born in England and a resident of the United States since 1951, Dyson has contributed significantly to the literature of theoretical physics. (He is, for instance, one of the four authors of the theory of quantum electrodynamics, and is said to have been excluded from winning the 1965 Nobel prize in physics only because the prize is limited to three recipients.) He is also a man of wide interests, ranging from interstellar spaceflight to technology policy, global ecology and science education. His writing on these subjects is clear, incisive and free from rhetorical excess.

I can think of no comparable author whose work is more selfless and genuinely helpful. *From Eros to Gaia* may be a collection of old stuff, but it adds up to an exceptionally rewarding read.

Emblematic of Dyson's poise and independence of thought are his essays here on the issue of how much money governments should spend on 'big science' endeavours such as the Hubble Space Telescope rather than more modest projects such as the International Ultraviolet Explorer. The Hubble Telescope cost a billion dollars. Its myopic main mirror made it an instant and embarrassing failure. The International Ultraviolet Explorer was launched in 1978, by a Delta rocket, on a shoestring budget. It has been pouring out data on young stars, supernovae, quasars and other celestial wonders ever since, sparking, Dyson notes, "more than a thousand papers in professional astronomical journals."

The lesson Dyson draws from all this is not the predictable one that big science is bad and small science good. In-

stead, he urges that attention be paid to the question of whether a big project can be accommodated within the "ecology" of the field to which it belongs. In particular, he sagely comments, planners should try to leave room for the waste and inefficiency required if scientists and engineers are to find out what works and what does not:

The right size means a size at which you can afford to take a gamble.... If you are too big to gamble, you are too big to do creative science.... The moral is not 'small is beautiful' [but] 'different is beautiful'.

Always one of our most visionary speculators about the far future, Dyson muses imaginatively about the long-term potential of spaceflight:

Spacecraft should weigh pounds rather than tons, they should cost tens of thousands rather than tens of millions of dollars, and they should fly missions at a rate of several per day rather than several per year.

He foresees an unmanned space probe for the year 2018:

I call this model the Astrochicken because it is about as big as a chicken and about as smart. It is a product of genetic engineering. It does not look like a chicken. It looks more like a butterfly. It has wide and thin solar sails instead of wings, and a high-resolution spectroscopic imaging sys-



**Machinery of the heavens** — a scholar, identified by his cap and gown, pierces the shell of the stars to observe the great wheels that move the heavenly bodies. The woodcut illustrates the mechanistic explanation of the cosmology proposed by the German cardinal Nicholas de Cusa (1401–1464). This picture is one of over a hundred reproduced in Harry Robin's *The Scientific Image: From Caveman to Computer*. Abrams, \$39.95.



tem instead of eyes. With its solar sails it flies around the inner solar system as far as the main belt of asteroids. At any one time there will be hundreds of such birds flying, programmed to make specialized observations of Earth, Moon, Sun, planets, and asteroids as well as the heavens beyond. Other cousins of the Astrochicken will have legs for landing and hopping around on asteroids, or solar-powered ion-jet engines for exploring the outer solar systems as far as Pluto.

Dyson rehearses his provocative notion that the countless comets of the Oort cloud thought to surround the Solar System are fit for human habitation. Comets, he suggests,

provide an open frontier, a place to hide and to disappear without trace, beyond the reach of snooping policemen and bureaucrats. ... Near to the Sun, space will belong to big governments and computerized industries. Outside, the open frontier will beckon as it has beckoned before, to persecuted minorities escaping from oppression, to religious fanatics escaping from their neighbors, to recalcitrant teenagers escaping from their parents, to lovers of solitude escaping from crowds.

Coming down to Earth, Dyson offers an analysis of carbon dioxide in the atmosphere that is sufficiently clear-headed and free from bombast to make it recommended reading for all parties in the global-warming debate, whatever their level of expertise: "In this chapter I shall be mainly concerned with facts and figures", he writes, and the facts and figures he provides are enlightening. Earth's atmosphere contains about 7,000 gigatons of carbon. The industrialized world is dumping more than six gigatons of carbon into the atmosphere every year. Deforestation and soil erosion add about another three gigatons annually. That's a total contribution of nine gigatons — yet measurements of the atmosphere show an increase of only half that much. Where is the missing carbon? Not in the oceans: "Oceanic measurements do not support" this interpretation. "So the mystery remains", Dyson notes. He concludes, not with an easy denunciation of the excesses of the developed world, but with a modest plea for "more observations".

With similar aplomb, Dyson calls for global monitoring of carbon's complement, the oxygen content of the atmosphere. "I am surprised that the general public all over the world is not clamoring to know how fast we are using up the oxygen", he writes. But he adds that it "is not only bureaucrats who are to blame. We scientists are also to blame for not practicing what we preach. It is easy to give advice, and not so easy to follow it." In a footnote he informs us that soon after his essay was published, in 1990, a young Princeton physicist took

up the challenge of making continual measurements of global oxygen levels. Were ecologists' contributions to the global-warming debate habitually couched in similarly thoughtful and constructive terms, they might more often see such prompt and palpable results.

Turning his attention to the classroom, Dyson offers the puckish precept that schools might in some cases do better to teach *less* science and more Greek and Latin — to function, in other words, as (classically) better schools. In an admirably balanced passage, he portrays the edifice of science as a hexagon with three ugly faces — science as "a rigid and authoritarian discipline, tied to mercenary and utilitarian ends, and tainted by its association with weapons of mass murder" — and three beautiful faces — "science as subversion of authority, science as an art form, and science as an international club". "The way to attract young people into science", he writes, "is to show them all six faces and give them freedom to explore the beautiful and the ugly as they please."

Education provides an insight into

Dyson's character. Never content to present himself as an oracle or sage, he seems always to keep learning, and he encourages his colleagues to do the same. At one point he recalls that he

shared an office for a month with a professor of sanitary engineering who taught me about biological oxygen demand and activated sludge. This was great fun, and my office mate was brighter than most of the physicists I know. I would advise any physicists who have a genuine concern for the environment to take the time to find out what the problems are in the field of activated sludge.

Seldom have the more attractive faces of science shone more brilliantly than on the pages of this estimable book. Dyson's writing casts a chalk-white beacon into many an out-of-the-way cranny, and illuminates all that he surveys. □

*Timothy Ferris is at the Graduate School of Journalism, University of California, Berkeley, California 94720, USA, and is the author of Coming of Age in the Milky Way, The Mind's Sky and other books on astronomy and physics.*

## Differing world views

E. G. Nisbet

**The Age of the Earth.** By G. Brent Dalrymple. *Stanford University Press*: 1991. Pp. 474. \$49.50.

**The Facts of Life: Shattering the Myth of Darwinism.** By Richard Milton. *Fourth Estate*: 1992. Pp. 288. £16.99.

KELVIN paid up. The sum was five shillings, and the bet was about the properties of radium. Would radioactivity invalidate his calculations of the age of the Earth? The winner was R. J. Strutt, later to become Lord Rayleigh; the Earth did indeed seem to be much older than Kelvin's estimate based on cooling. In 1904, Rutherford lectured on the topic. He described his apprehension when he spotted Kelvin in the audience; to his relief, Kelvin fell asleep, but as Rutherford came to the critical point, he saw the "old bird" sit up, open an eye and cock a baleful glance. Rutherford escaped by pointing to the heat generated by radioactivity, and the old boy beamed at him.

Since Kelvin's day, it has been unanimously accepted by orthodox scientists that the Earth is old. Radioactive heat was not the only reason Kelvin was wrong: the Earth's interior is convective, not conductive, and is still slowly losing both *primaeval* and radiogenic heat. The Earth is roughly 4.5 billion ( $10^9$ ) years old. Although there is still debate about

the accuracy and precision of that number, its general magnitude is certain. If the number is very seriously wrong, and the planet is only a few thousand years old, then not only geology but a good chunk of physics, chemistry and applied mathematics is also in error. Rutherford realized the importance of his discoveries. On his Nobel citation (a lovely document — there is a copy at McGill University) are the classic curves of radioactive decay, and he chose them for his coat of arms. They are central to much of modern physics.

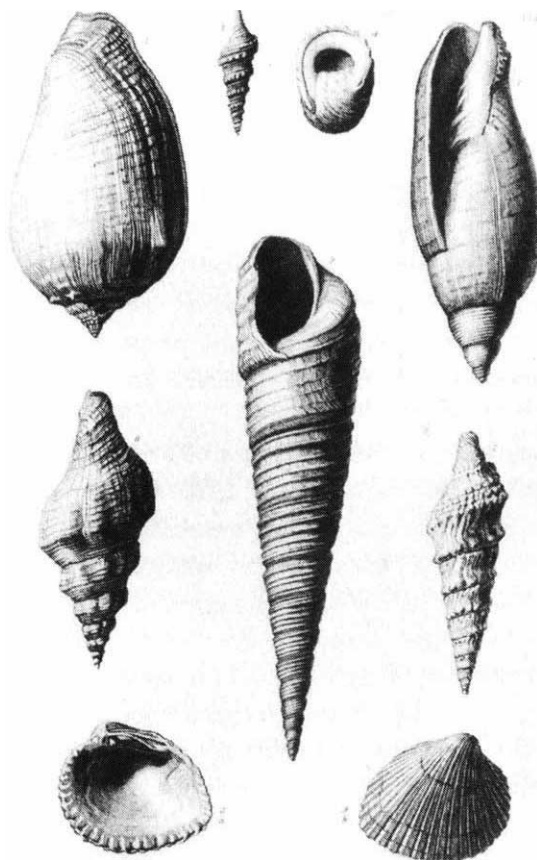
G. Brent Dalrymple is a widely respected geochronologist, and his book is a fine exposition of the enormous body of evidence that shows, unequivocally, that Rutherford was right, the Earth is old. Ironically, we have the creationists to thank for this book, which was written in answer to their onslaught. Dalrymple's scholarly but readable account passes from an opening act on the early debates about the planet's age, to the main argument on the basis of radiometric dating. Once the groundwork is done, the book progresses through a discussion of Earth's oldest rocks and the direct evidence for old crust, to chapters on the supporting cast — the Moon and meteorites — and then gets down to the hard action of lead isotopes, followed by a finale on the Universe and all that it contains. Thorough-

out the work, the writing is careful, un-journalistic but well written science. Parts of it are not especially easy — lead isotopes in particular are horrendously complex. Putting an exact number on the age of the Earth is difficult — Dalrymple favours a best value around 4.54 billion years, but the error is around 0.02 billion years. Rutherford's basic point is firm: the Earth is not thousands but billions of years old.

Richard Milton does not accept this, nor does he accept a lot of the rest of the conventional wisdom of biology and geology. Most authors who attack the fabric of geology are creationists, mostly fundamentalist Christian but some Islamic; Milton, in contrast, is a secular writer who states that he holds no religious belief. His work is a work of journalism, in the British tradition that puts opinion on the front page, giving only one side of the debate and placing the judgement of the writer foremost. Consequently, the book has the opinion-loaded theological style of the English 'quality' newspapers (tell it to the cobelievers), not the quiet accuracy of the *New York Times*. The result is an entertaining but wildly misleading work, more in the ways of F. R. Leavis than T. H. Huxley (no stranger to opinion himself), that wholly misunderstands the basis of stratigraphy. Milton's book dismisses the painstakingly careful logic of radiometric dating in a few pages, in contrast to Dalrymple's thorough discussion of lead isotopes. What the book does say about isotopes, for instance about radiogenic helium, is hilarious or irrelevant, and the book evades the rigour of Dalrymple's evidence by the simple device of ignoring it. At least the fundamentalists, on occasion, attempt to meet geochronology squarely; not so here, as the work leaves out virtually all the facts.

Why then bother to review a book that is so flawed? Although Milton spends many pages attacking the view that the Earth is old, the book's main target is not physics and chemistry but biology. The cover has a picture of a shattered Darwin, and with a passing reference to Marxism, the book attempts to consign Charles to the same dustbin as Karl. Here there is a valid point: "ideological Darwinism has replaced scientific Darwinism in our educational system". The Nazis misused Darwin, and so did the Soviets, and the result was inhumanity on a grand scale. In democratic societies too, the impact of the eugenicists was powerful in the first half of the century, and their ingrained influence remains. Even today, the detestation of the 'inferior' is widespread, for instance in the medical profession's commonly held opinion that 'defective' children should not be born. More generally, Tennyson's view of "Nature red in

**Fossil mollusc shells characteristic of Charles Lyell's 'Miocene' period, based on the work of the French geologist Deshayes. The drawing originally appeared in volume three of Lyell's *Principles of Geology* (1830–1833), a facsimile edition of which was reviewed by Stephen Jay Gould in *Nature* 355, 215 (1992); it is reproduced here from *The Discovery of Evolution* by David Young, a lavishly illustrated account of the history of evolutionary theory. Cambridge University Press, £40 (hbk), £14.95 (pbk).**



tooth and claw" is used to justify the excesses of 1980s capitalism. But are these valid reasons for attacking Darwin himself, who knew none of this?

More vulnerable to attack are some of the great panjandrums of palaeontology. Their discoveries have been enormous, both intellectually and (sometimes) physically (especially the dinosaurs). The best of them are scientists in the classic sense: humble searchers after truth. But publicity and fame often go to the less pure. Among these, the feuds, arrogance and frequent intellectual dishonesty are legendary. One thinks, for instance, of the forger of Piltdown man, of some of the dinosaur discoverers, or of Marie Stopes, whose status as a foundress of palaeobotany and contraception went along with disturbing views on eugenics. But palaeontologists are human too; many have been monuments of integrity, and in their defence is the record that they are their own best critics. As in many things in human history, one should be careful to separate the discovery itself (if true) from the people and motives that led to its uncovering.

Evolution *has* occurred; the Earth is old; Milton's book is simply wrong. Nevertheless, the book has a point buried deeply in its farrago of misfact. Evolution (especially in bacteria) is a much more subtle process on a molecular level than Darwin could have im-

agined, and we should be careful to reflect this in our teaching. Indeed, if we do fall into the trap of comparing human society with nature, then it is the information exchange between bacteria that we should cite, not the record of eukaryote 'survival of the fittest'. There is still much we do not understand about the controls on evolution. Even isotopic systematics have their continuing interest, as hordes of geochemists can testify, and we are still far from understanding the complexities of geochemical cycling.

With luck, someone will read Dalrymple's careful honest work, take note, and be well guided about the truth of Earth history, as far as we can decipher it. *The Facts of Life*, I fear, will make a greater impression, in misleading its readers: it is worth recalling the darwinian comment that, while incorrect hypotheses are the life of science, incorrect 'facts' are its death. Nevertheless, we can take refuge in the classical Miltonic view of the Areopagitica that the facts are too strong to be obscured by misinformation. Perhaps if it teaches us some humility, then even a bad book can do a little good. □

E. G. Nisbet is in the Department of Geology, Royal Holloway College, University of London, Egham, Surrey TW20 0EX, UK.



# Geological granddaddy

Martin Rudwick

**James Hutton and the History of Geology.** By Dennis R. Dean. Cornell University Press: 1992. Pp. 303. \$38.50.

SCIENTISTS often crave a founding father-figure for their science; their motives would make an instructive study in the anthropology of science. The choice of patriarch is often chauvinistic, revealing modern science to be as internationalist as the Olympics. For geologists, the British (indeed English-speaking) favourite has been James Hutton (1727–97), the Scottish author of a massive *Theory of the Earth* (1795) that is far more often revered than read. A quarter of a century ago, E. B. Bailey, a former director for the British Geological Survey, expressed the claim baldly with his book *James Hutton: The Founder of Modern Geology* (1967). The title of Dennis Dean's new book is more ambivalent, but he ends his text by endorsing Bailey's phrase. Meanwhile, most historians of science have grown reluctant to identify either founders or forerunners, judging that both epithets obfuscate more than they reveal.

Hutton is a difficult subject for biography, because few manuscript records by or about him are now extant, and virtually all his scientific work was published in the last few years of his life. Nonetheless, Dean has done a fine job of piecing together what he and other historians have managed to ferret out about Hutton's career. It was a remarkably uneventful one in outward terms. Hutton was a typical figure of the European Enlightenment, and a prominent one in its Scottish manifestation — a friend of David Hume and Adam Smith. After a standard medical education, much of it on the Continent, he was for some years actively involved in agricultural improvement, until a sleeping partnership in a successful chemical firm gave him the financial freedom to settle in Edinburgh as a somewhat eccentric bachelor and a leisured 'natural philosopher'. That contemporary term, rather than the anachronistic word 'scientist', defines his intellectual endeavour: he would have regarded his three-volume *Principles of Knowledge* (1794) as just as important as his geological *Theory*.

The most valuable aspect of Dean's book is that it systematically summarizes all Hutton's writings about the Earth; his prolix texts are expounded with clarity and insight. This gives readers a reliable summary of what Hutton claimed, and on what kinds of empirical and concep-

tual foundations; it should act as a helpful guide in reading the texts themselves. The same expository treatment is extended to the immediate reception in Britain and Ireland of Hutton's great paper to the Royal Society of Edinburgh (1785) and to that of the two-volume *Theory* that set out his argument in detail a decade later (a third volume remained unpublished at his death, and indeed until 1899). The exposition also covers the work of Hutton's younger champion, the Edinburgh mathematician John Playfair, whose *Illustrations of the Huttonian Theory of the Earth* (1802) was the more readable, but significantly



**Hutton — eccentric bachelor and leisured natural philosopher.**

modified, medium by which the majority of nineteenth-century geologists imbibed their Hutton. Although Dean focuses on Hutton's texts and those of his British and Irish contemporaries, he also reproduces a good selection of the drawings with which they illustrated their arguments, and he is sensitive to the significance of their concrete first-hand experience in the field.

The book would have been more coherent and even more valuable, however, if it had been restricted to Hutton's lifetime and the first decade or two after his death. As it is, Dean pursues an increasingly diluted stream of 'Huttonian' geology, which he concedes was largely mediated through Playfair, all the way to Charles Lyell's avowedly neo-Huttonian *Principles of Geology* (1830–33), and even through to the end of the century. This later part of the book is increasingly sketchy, and becomes largely derivative of the work of other historians. It ends with a

rapid survey of "Hutton and the historians"; but this turns out rather curiously to mean various scientist-historians of the nineteenth century, and the survey stops around 1900 with the venerable Geikie and Zittel. What is therefore missing from Dean's concluding assessment of Hutton's place in the history of geology is a thorough engagement with the research of modern historians such as Roy Porter (*The Making of Geology*, Cambridge University Press, 1977) and Rachel Laudan (*From Mineralogy to Geology*, University of Chicago Press, 1987), both of whom see Hutton as a figure somewhat apart from the mainstream of European or even of British geology, and for whom the controversies between Huttonians and their critics were by no means a matter of black and white. Such an assessment does not make Hutton any less important, of course, but it does demand at least a reconsideration of the traditional interpretation of Hutton as the, or at least a, 'founder' of geology.

Dean's book would also have been more useful if it had provided a richer context for Hutton's work. Dean claims, in my opinion correctly, that Hutton's famous conclusion, that in the Earth "we find no vestige of a beginning, no prospect of an end", should be read primarily as a statement about the limitations of human knowledge, rather than as a claim about the vast timescale or even the eternity of the world. But that interpretation needs to be related to Hutton's own epistemology; Dean has missed the chance to present Hutton's intellectual enterprise as a unified whole. Furthermore, even Hutton's 'geology' needs to be understood in his own fully international context. Like other Enlightenment philosophers, Hutton was well aware of what was going on in countries other than his own; he thought nothing of including in his book lengthy untranslated quotations (mainly from the great Genevan naturalist Horace-Bénédict de Saussure) in the then international language of science. Unfortunately Dean gives only the most cursory attention to Hutton's Continental sources, or to the way Hutton's theory was received beyond the English Channel. Above all, he does not show how Hutton's work fitted into a well established and pan-European tradition of 'theories of the Earth', nor does he analyse adequately how Hutton's book and others in the same genre differed from the self-consciously new science of geology that replaced them in the decades after Hutton's death. □

*Martin Rudwick is in the Department of History and in the Science Studies Program, University of California, San Diego, La Jolla, California 92093-0104, USA.*

# Closing the hominid gap

Glenn C. Conroy

**Origins Reconsidered: In Search of What Makes Us Human.** By Richard Leakey and Roger Lewin. Doubleday/Little, Brown: 1992. Pp. 375. \$25, £18.99.

In the hands of skilled storytellers, the international race to close the 'hominid fossil gap' has all the excitement, romance and intrigue of a modern-day Manhattan Project — the only difference is that human-origins research is often portrayed as being more explosive. In

phylogenetic tree, is the level around which most verbal fisticuffs in palaeo-anthropology arise. For the most part, Leakey and Lewin spare the reader any dogmatic pronouncements on the first level of analysis and quickly get to the more interesting, and of course more

controversial, second and third levels. As a refreshing antidote to the metastases of cladistic theology in palaeoanthropology, they place more emphasis on how these discoveries affect our notions about human evolutionary grades and the biological adaptations shown by each (for example, australopithecine versus *Homo erectus* grades of adaptation). This is the most enjoyable, informative and illuminating aspect of the book (for example, the discussion of maturational patterns in Turkana Boy). To a large extent, the authors avoid the typical preoccupation of many palaeoanthropologists with defining direct fossil lineages (although Leakey's understandably Kenya-

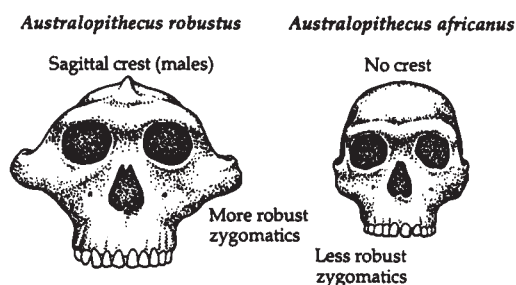
centric view of human evolution does peek through at times). When one considers that all known fossil hominoids from Africa come from only a tiny fraction of the present-day land surface of that continent, it stretches credibility to believe that any direct ancestors of later humans have yet been found. The recent discovery of a Miocene hominoid from northern Namibia is a powerful reminder of how little we know about the geographic distribution of human ancestors.

To be sure, *Origins Reconsidered* has many of the elements the public now comes to expect from trade books on human origins. There are the usual tales of adventure in remote and dusty places and the heavy dose of male bonding (or 'male bondage' as Dan Quayle once inadvertently called it). Palaeoanthropologists have elevated the 'don't get mad, get even' school of scientific journalism to an art form and the reader does not have to wait long before Leakey and colleague Alan Walker return the favour by effectively 'slam-dunking' Donald Johanson. Leakey teases his audience a bit by lamenting that his personal and professional relationship with Johanson began to deteriorate for unspecified reasons he considers "best not discussed

publicly". The audience is thus left to its own imagination: let's see — espionage? ivory smuggling? polyester safari suits? ordering red wine with fish?

While the narrative of the book is a relatively straightforward account of discovery and interpretation, a certain amount of poetic licence creeps into the story. For example, molecular biologists would be surprised to learn, and David Pilbeam would blush to read, that it was Pilbeam's "shift in position" that legitimized the field of molecular anthropology. Similarly, Phillip Tobias might shake his head to read that Raymond Dart "unearthed many hominid fossils between the 1930s and 1950s from four main cave sites in South Africa". As far as I am aware, until the fossils discovered at Makapansgat in the late 1940s by Tobias and colleagues rekindled his interest in palaeoanthropology, Dart displayed little personal initiative in exploring the cave sites of South Africa (including Taung). Luckily, Robert Broom was not so deterred. In fact, one might draw the analogy that Robert Broom was to Dart what Thomas Huxley was to Darwin.

A number of current debates are reviewed, including the molecular evidence for a human-chimpanzee clade and the 'mitochondrial Eve' story. If it is true, as some molecular anthropologists argue, that two such morphologically

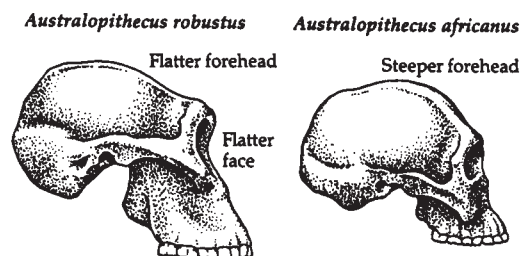


Differences between these two species...

their comfortably chatty and familiar style, Richard Leakey and Roger Lewin use the recent discoveries of a nearly complete *Homo erectus* skeleton (the 'Turkana Boy') and a hyper-robust australopithecine cranium (the 'Black Skull') as convenient launching pads for this sequel to their successful previous collaboration, *Origins*, which was published in 1977. Leakey's unique personal history combined with Lewin's crisp, engaging writing style results in an enjoyable, fast-paced read.

Leakey is described on the dust jacket as "the world's most famous living paleoanthropologist", leaving me to ponder whom the publisher's regard as the world's most famous dead palaeoanthropologist. The only disappointing aspect of the book's production is the poor choice and marginal quality of the photographs. Inexplicably, there is only one (inadequate) photo each of the Turkana Boy and the Black Skull, even though they are the centrepieces of the story.

Both discoveries provide important new insights into human evolution, particularly the *Homo erectus* skeleton. Most evolutionary hypotheses can be broken down into three analytical levels: the cladogram, the phylogenetic tree and the scenario. Unlike cladograms and phylogenetic trees, scenarios are not diagrams but, rather, historical narratives that describe not only phylogenetic relationships but also the ecological and evolutionary forces that most directly influence the organisms under discussion. The scenario, being more hypothetical and interpretive than the cladogram or



...lie mainly in dental and cranial features.

dissimilar animals as humans and chimpanzees share a more recent common ancestor than two such morphologically similar animals as chimpanzees and gorillas, then palaeontologists have little choice but to pack up their shovels, throw in the towel, and never again attempt another cladogram based on morphological characters. But do molecular anthropologists overstate their case? Vince Sarich once tweaked the collective noses of palaeoanthropologists by writing that "one no longer has the option of considering a fossil specimen older than about eight million years a hominid no matter what it looks like". However, let us remember that the human genome contains some three billion base pairs, so that even 'good' DNA sequence data of 10 kilobase pairs cover only a small fraction of the genome. Given the recent 'mitochondrial Eve'



fiasco in which flaws in statistical procedures and maximum parsimony analysis of the mitochondrial DNA data using the PAUP computer program have been noted and the conflicting molecular evidence about whether the chimpanzee-gorilla-human trichotomy is really resolved, or even resolvable, one 'genioglossus in vestibule' rejoinder to Sarich might be that "one no longer has the option of considering chimpanzees more closely related to humans than to gorillas *no matter how good the molecular evidence superficially appears to be*".

To his credit, Leakey is not hung up on giving fossils specific names and seems to understand, unlike many of his peers, that just because one gives a specific name to a fossil does not necessarily mean that one has described a real biological species. The biological species concept is only operative in organisms where reproductive behaviour can be examined in nature (or the laboratory). Species-specific morphological distinctions can only be codified after species are so identified. The process does not work the other way round. One cannot first observe differences in fossils and then claim that these differences define true biological species because, as is well known, morphological differentiation and speciation are not necessarily correlated. One reason evolutionary scenarios are so contentious is that palaeoanthropologists continually name fossil species and then create scenarios based on the artificial constructs they have just created. So, for example, the discovery of the Black Skull may indeed force people to "change their minds", as Leakey and Lewin put it, about australopithecine phylogeny, but only those who bought into the 'orthogenetic' view of australopithecines as evolving linearly from *A. africanus* through *A. robustus* to *A. boisei* in the first place. For those of us who noted a decade ago that this scheme was unlikely, our views are now reinforced, not changed.

As a self-made man, Leakey has had to put up with a lot from his detractors, some of whom would undoubtedly concur with Disraeli that self-made men often tend to worship their creator. But I, for one, have a great deal of admiration for the man. Part of the reason is that, like me, he studies fossils the old-fashioned way — he tries to find them. But it is his passion for the African bush and the preservation of its vanishing wildlife that resonates most strongly with me. In this, the latest and most difficult challenge of his life, we all must see to it that he succeeds. □

Glenn C. Conroy is in the Departments of Anatomy and Neurobiology, and of Anthropology, Washington University Medical School, St Louis, Missouri 63110, USA.

## Common-sense consciousness

Stuart Sutherland

**The Rediscovery of the Mind.** By John R. Searle. MIT Press: 1992. Pp. 270. \$22.50, £19.95 (hbk); £9.95 (pbk). (Pbk not yet available in the US.)

JUST as the mini-skirt comes and goes in women's clothing, so does consciousness fluctuate in fashion for philosophers. In the past few years, bookshops have been graced or despoiled by shelf-loads of tomes devoted to the subject. Almost all have been turgid and obscurantist. Now, with *The Rediscovery of the Mind* we have a clear, well written and cogently argued account, which is marked by that unfashionable quality, common sense. In standing up for the existence of consciousness as a fact, John Searle attacks with pleasing acerbity the loonier beliefs of those who have attempted to deny it. For example, of the Churchlands, who claim that mental terms have been invented by ordinary people to explain behaviour ("folk psychology") and therefore have no validity, he asserts: "It is hard to come right out and say that no one ... ever drank because she was thirsty ... but it's easy to challenge something if you can label it in advance as 'folk psychology'." And on other efforts to eliminate consciousness by regarding it as identical with a high-level account of certain brain operations, he writes, "If you are tempted to functionalism, I believe you do not need refutation, you need help".

Searle's own view is that "mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain" and he rightly insists on the distinction between consciousness and behaviour or even, as Gilbert Ryle put it, dispositions to behave, for "a system could have consciousness without behaviour and behaviour without consciousness." All this is robust common sense and, I believe, unarguable.

Searle runs into trouble, however, when he attempts to specify the relation between consciousness and the brain. He writes, "Consciousness is a higher-level or emergent property of the brain in the

utterly harmless sense of 'higher-level' or 'emergent' in which solidity is a higher-level emergent property of H<sub>2</sub>O molecules when they are in a lattice structure (ice)", but he goes on to deny the reducibility of consciousness to brain function. But if solidity can be reduced to the behaviour of molecules, why cannot consciousness similarly be reduced to the behaviour of the brain of which it is an emergent property? Searle's answer is tortuous. He claims that solidity originally meant the feeling produced by a solid object: the reduction applies not to this feeling but to the property of the object that gives rise to the feeling. This argument does not work: if the solidity of an object can be reduced to the behaviour of its molecules and if the relation between consciousness and the brain is analogous to that between the solidity of an object and the state of its molecules, then surely consciousness should be reducible to brain states. The contention that consciousness is merely an emergent property of brain states cannot therefore be sustained. It is no use Searle arguing, as he does, that such phenomena as magnetism were irreducible before Maxwell and similarly that one day we will have a theory that shows how consciousness emerges from the brain. There must be, and indeed is, something special about consciousness if it is not reducible.

Searle is also weak on the evolution and function of consciousness. He thinks it allows us to make finer discriminations and has no hesitation in assigning to it a causal role in behaviour. But elsewhere he talks of two robots that behave identically, one of which is conscious (because



**Lizard (1959) by M. C. Escher, one of more than four hundred illustrations reproduced in Doris Schattschneider's beautiful *M. C. Escher: Visions of Symmetry*. The book is now published in paperback by W. H. Freeman, £19.95, \$24.95. For a review, see *Nature* 349, 471 (1991).**

we have designed into it whatever properties of the nervous system give rise to consciousness) and the other is not. He cannot have it both ways — either consciousness has a function or it does not.

Searle is interesting and sound on intentionality, that is, the way in which all cognitive processes refer to entities. Underlying any belief or proposition are a vast number of other beliefs and unacknowledged assumptions. You cannot interpret the command "Sit on the chair" without the assumptions (perhaps never consciously made) that the seat is solid and that solid surfaces can support the human bottom. From his discussion of intentionality, he proceeds to attack the notion that there are unconscious thoughts that cannot in principle be made conscious. He would allow that a memory is mental even if it is not currently in consciousness, but he denies that there could be unconscious ideas that could never be brought to awareness, attacking in the process parts of Freudian theory. Although he does not mention him by name, he dismisses Helmholtz's idea of unconscious inference, as exhibited, for example, in our remarkable capacity to see depth using at least ten different cues.

Searle argues that the neural activity that results in conscious experience must be specified at a neurophysiological level and that it is vacuous to regard the brain as a digital computer, because everything from whirlwinds to leaves can be regarded as computers depending on the observer's point of view. His attack on cognitive science seems misguided. It is of course true that computer simulation may help us to understand the behaviour of whirlwinds. From a scientific point of view it is irrelevant whether the hypothetical variables postulated in attempts to understand brain function (such as David Marr's primal sketch and 2½-dimensional sketch) are or are not mental, provided they help us to understand at a high level of abstraction how the brain is working.

It is a pity that Searle ends with this sort of logic chopping, particularly as he gives no advice on what cognitive scientists should be doing other than what they are doing, except for an injunction that they should be more cautious in their claims and in their use of language. Indeed it is doubly a pity, for the last chapters mar a book that is almost always sensible and lucid, so much so indeed that it is likely to be ignored by that vast majority of potential readers who believe books on consciousness are only of interest if they make unintelligible departures from common sense. □

Stuart Sutherland is in the Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, UK.

## Brian's brain

Igor Aleksander

**The Turing Option.** By Harry Harrison and Marvin Minsky. Warner/Viking: 1992. Pp. 422. \$21.95, £14.99 (hbk); £8.99 (pbk).

MARVIN Minsky, a doyen of artificial intelligence (AI), may be remembered by some for his prediction, made in the early 1970s, that three to eight years from that time we would have a machine with the general intelligence of a human being. I have often wondered how he now feels about this prediction, which has evidently not come to pass. How has the wisdom accrued over 20 years of further thought dealt with this aberration of judgement? Perhaps one way of dealing with it is to make the prediction come true through a novel. *The Turing Option*, which Marvin Minsky has written with Harry Harrison of *Make Room! Make Room!* fame (the film version may be remembered as *Soylent Green*), is the story of the ultimate AI — how things might be in the 2020s had Minsky's prediction been true.

Brian Delaney, a brilliant Irishman, works in the United States for Megalobe, where he has just invented the "real" AI. But as he is about to demonstrate it, the megacooks get into his lab, put a bullet in his head and steal the AI. Despite the bullet, Brian is not dead. The equally brilliant Dr Snaresbrook is called in to rebuild his brain, which she does using the latest micromanipulative technology, supercomputer implants and a deep knowledge of the way that a mind can be rebuilt by knowing a lot about the way it is related to the neural structure of the brain. It's not telling tales out of school to say that Brian not only recovers, but retraces his inventive steps and redesigns the AI. Will Brian and the AI catch the crooks? Will they develop AI products, become millionaires and live happily ever after? Oh, and one more thing about the plot, the AI becomes a tree-like robot with human charms and social skills. It objects to being called AI and becomes MI, Machine Intelligence ("there's nothing artificial about me").

My personal problem with science fiction of this kind is that knowing that one of the authors is an eminent scientist makes me aware of two plots, one scientific and the other literary. Even worse, it makes me expect the first of these to be at least scientifically interesting, if not always plausible. Unfortunately, the interesting scientific story gets dreadfully polluted with high-tech gimmickry. From the first few pages, the authors shower their readers with voice-sensing telephones, high-definition television con-

ferencing, voice faxes and a host of other minor gadgets. No problem here with scientific plausibility, it's just the yawns that have to be suppressed. But digging past this high-tech veneer, the scientific plot touches on interesting points. Can the mind of a damaged brain be reconstructed by reconnecting the severed connections? Can the process be aided by implanting a microminiature computer in the brain? Brian discovers to his surgeon's delight and surprise that the computer in his brain not only serves to control some mental activity, but can also be accessed, and its memory contents become part of Brian's consciousness — that's definitely food for thought. Then there is Sven, or Machine Intelligence. He is given masses of silicon memory and that seems to be sufficient for him to become increasingly 'human' as the plot progresses. Is bulk memory what human intelligence is all about?

These debating points are probably the best there is about this novel. Unfortunately, the literary plot failed to capture my interest. Again it's the choking effect of high-technology that may be to blame. Sven the robot takes centre stage among the characters. It seems to have attracted all of the authors' capacity for characterization, leaving Brian and his human friends as rather wooden figures that have little in the way of humanity. Giving the benefit of the doubt, this may be part of the design. It makes the point that machines could be more human than humans. This may be subtle, but takes away from the pace and tension that might be expected from what the dust jacket promises will be "a thriller of hair-raising suspense".

There may well be readers who are thrilled by the thought of voice faxes and supersmart robots in a world where people sink into the background and are only worthwhile if they are capable of designing intelligent machines or performing extraordinary feats of mind surgery. My own literary needs for thrills and suspense are better served by the likes of John Le Carré or Len Deighton. □

Igor Aleksander is in the Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, UK.

■ A new edition of *What Computers Still Can't Do* by Hubert L. Dreyfus has just been published by MIT Press. In 1972, this manifesto on the inherent inability of machines to mimic higher mental functions caused an uproar in the AI community. Dreyfus has now added a lengthy introduction addressing the shift in the focus of research towards connectionism and neural networks. Price \$27.50, £24.95 (hbk); \$13.95, £12.50 (pbk).



**Pick of the paperbacks**

■ *Chemical Evolution: Origins of the Elements, Molecules and Living Systems* by Stephen F. Mason, Oxford University Press, £8.95. For a review see *Nature* **351**, 199 (1991).

■ *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*, second edition, by Lynn Margulis. W. H. Freeman, £27.95.

■ *Fractals, Chaos, Power Laws* by Manfred Schroeder, W. H. Freeman, \$19.95, £15.95. For a review see *Nature* **350**, 524 (1991).

■ *The Great Mambo Chicken and the Transhuman Condition: Science Slightly Over the Edge* by Ed Regis. Penguin, £6.99. "Virtuoso science writing of rare vigour and mastery", wrote Walter Gratzer in a review in *Nature* **354**, 338 (1991).

■ *The Character of Physical Law* by Richard Feynman, a series of lectures given at Cornell University in the mid-1960s. Penguin, £5.99 (with an introduction by Paul Davies); MIT Press, £5.95, \$8.95.

■ *Another Fine Math You've Got Me Into* by Ian Stewart, a collection of mathematical puzzles. W. H. Freeman, \$13.95, £10.95.

■ *The Changing Atmosphere: A Global Challenge* by John Firor. Yale University Press, \$9, £5. For a review see *Nature* **349**, 472 (1991).

■ *Discovering Egyptian Hieroglyphs: A Practical Guide* by Karl-Theodor Zauzich. Thames and Hudson, £7.95.

■ *Sell Yourself To Science: The Complete Guide to Selling Your Organs, Body Fluids, Bodily Functions and Being a Human Guinea Pig* by Jim Hogshire. Loompanics, Washington, \$16.95.

## A book for burning

Martin Gardner

**Spontaneous Human Combustion.** By Jenny Randles and Peter Hough. *Robert Hale*: 1992. Pp. 224. £14.95.

HERE we go again — another worthless book pandering to public obsession with the paranormal. The authors, Jenny Randles and Peter Hough, have separately and in tandem perpetrated previous books about UFOs, crop circles, witchcraft and similar wonders. In this book they do their best to persuade readers that there have been hundreds, perhaps thousands, of cases of people bursting into flames from causes science cannot fathom.

The simplest explanation of most such burnings is that victims fall asleep, often intoxicated, and set fire to their clothes with a cigarette, cigar or pipe. If elderly, fat and dehydrated, and given sufficient time, almost all of the body can be reduced to ashes. Because the media love to sensationalize, sometimes to lie, the myth of spontaneous human combustion continues to thrive.

Not a single chemist or physicist of repute is cited in this shameful book, and for an obvious reason. None believes that a human body can spontaneously catch fire. And virtually all the periodicals cited in the book are not scientific journals but obscure fringe magazines devoted to the paranormal.

The authors discuss a variety of ways in which a body might ignite apart from contact with such things as burning cigarettes and fireplaces. Perhaps ball lightning enters a room to smite the

victim. Static electricity? (The authors call this a "shocking suggestion".) Force-fields from high tension wires? Psychic curses from enemies? Poltergeists? "Kundalini fire", an energy that Yogis claim pervades a person's astral body? One "researcher" thinks the body may contain a subatomic particle — he calls it a "pyrotron" — that undergoes a chain reaction unknown to physicists.

The authors take seriously what they call a "phosphinic fart". That expelled gas is highly inflammable is easily demonstrated. Could it be, the authors wonder, that such gas might float up through clothing to be fired by a cigarette?

The sole merits of this preposterous book are that it catalogues more than a hundred alleged cases of spontaneous human combustion, cites earlier books even worse than this one, and briefly considers cases of the phenomenon in fiction. The earliest example is in Charles Brockton Brown's novel *Wieland*. Others include Nicolai Gogol's *Dead Souls*, Captain Marryat's *Jacob Faithful*, Herman Melville's *Redburn* and Emile Zola's *Le Docteur Pascal*. The most famous instance is the spontaneous human combustion of a drunken Mr Krook in Charles Dickens's *Bleak House*. Dickens was soundly trounced by George Henry Lewes, George Eliot's live-in lover, for treating spontaneous human combustion as a fact. (See *Blackwood's Edinburgh Magazine*, volume 89, April 1861.) Their debate was notable among the many controversies about the subject that raged in nineteenth-century England.

The jacket of *Spontaneous Human Combustion* shows a grizzly photograph of the remains of Mary Reeser, a plump 67-year-old resident of St Petersburg, Florida, who burned to death in 1951.

The authors devote a chapter to this most famous of all modern cases. When Mrs Reeser caught fire, she had not eaten dinner, was wearing a flammable nightgown, had taken two Seconal tablets and was smoking a cigarette. When her son kissed her goodnight before leaving, she said she intended to take two more sleeping pills. Not until the following morning were her ashes found. The well-stuffed chair in which she sat was also consumed. The authors cannot imagine why the house did not burn down, ignoring the fact that the floor was concrete and a ceiling beam was burning when firemen arrived.

The Federal Bureau of Investigation's laboratory report on the case, based on many similar investigations, stated: "Once the body starts to burn there is enough fat and other inflammable substances to permit varying amounts of destruction to take place. Sometimes this destruction by burning will proceed to a degree which results in almost complete combustion of the body." Yet the authors strive valiantly to avoid the obvious and to convince readers that something strange and inexplicable had felled Mrs Reeser.

Why are cases of alleged spontaneous human combustion so much rarer today than in past centuries? Because reporting is more accurate and forensic methods have improved. In 1982 the US press had a field day when a Chicago woman was said to have burst into flames while walking along a street. When a medical examiner investigated, he found that the woman had been dead for 12 hours, and her clothes doused with gasoline.

On their last page the authors conclude that spontaneous human combustion is "triggered by forces that are just as puzzling to us as electricity was to Leonardo da Vinci. Yet, like electricity, it is probably a natural — not supernatural — energy dormant in our midst. If we can learn to tap into it, to harness its raw power and perhaps control its anger, then who knows what advantages it might bring to our children?"

Elsewhere, indulging in another pun, the authors call the 'mystery' a "burning question" for modern science. It is not a burning question. It is not even a question. Not a single textbook on forensic medicine written this century considers the phenomenon a possible cause of death. The only burning question is how today's cynical publishers can maintain a clear conscience when they flood the market with absurd potboilers solely for the purpose of extracting money from an ill-informed, gullible public. □

Martin Gardner is at 110 Glenbrook Drive, Hendersonville, North Carolina 28739, USA.